



A continuous time-and-state epidemic model fitted to ordinal categorical data observed on a lattice at discrete times

Rémi Crete, Besnik Pumo, Samuel Soubeyrand, Valérie Caffier, Frédérique Didelot

► To cite this version:

Rémi Crete, Besnik Pumo, Samuel Soubeyrand, Valérie Caffier, Frédérique Didelot. A continuous time-and-state epidemic model fitted to ordinal categorical data observed on a lattice at discrete times. 2012. hal-00735874

HAL Id: hal-00735874

<https://hal.science/hal-00735874>

Preprint submitted on 27 Sep 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A continuous time-and-state epidemic model fitted to ordinal categorical data observed on a lattice at discrete times

Auteurs. Rémi Crete^{1,2,*}, Besnik Pumo^{1,2}, Samuel Soubeyrand⁴, Valérie Caffier³, Frédérique Didelot³

¹ UMR LAREMA, Université d'Angers, 2 bd. Lavoisier, 49045 Angers, France

² Agrocampus Ouest - Centre d'Angers, 2 rue le Nôtre, 49000 Angers, France

³ INRA, Institut de Recherche en Horticulture et Semences (INRA, Agrocampus-Ouest, Université d'Angers), SFR 149 QUASAV, F-49071 Beaucozéz, France

⁴ INRA Avignon ... á compléter, Avignon, France

* e-mail : remi.crete@etud.univ-angers.fr

Abstract.

We consider a spatio-temporal ¹ model to describe the spread of apple scab within an orchard composed of several plots. The model is defined on a regular lattice and evolves in continuous time. Based on ordinal categorical data observed only at some discrete instants we adopt a continuous time approach and apply a Bayesian method for estimating unknown parameters.

1 Introduction

In this paper we propose a model for describing an epidemics of a plant disease evolving in space and time and study statistical inference of the unknown parameters. The scientific literature is very rich about spatio-temporal data and we refer the reader to the book of Cressie and Wikle (2011) for an exhaustive description of associated models. The biological process under consideration can be seen as spatio-temporal process evolving over a lattice and continuous in time, but the associated observations are ordinal categorical data sparsely spaced in time.

Considering time as continuous or discrete component leads to two different approaches (see Mollison (1977) for a general description of space-time models). The discrete time approach is used to analyse such data by Chadoeuf et al. (1992), Guyon and Pumo (2007), Zhu. et al. (2008) among others. Following the ideas of Gibson (1997), Rasmussen et al. (2007), Soubeyrand et al. (2009) etc. we adopt in this work a continuous time approach and propose a Bayesian inference for estimating the unknown parameters. Nevertheless in opposition to the above cited works where data are binary in our work we analyse ordered categorical for space-time data which is new to our knowledge.

The model tends to describe the dynamic of apple scab within an orchard (a spatial lattice) by different factors such as climatic conditions and different forms of pathogen. It is defined as a multivariate point process (Daley and Vere-Jones, 2003), with spatial dependence. The main purpose of this work is to understand and describe statistically the interaction between trees. We state conditions under which our model is uniquely determined and point out the possible improvements that permits to adapt the model in similar situations.

The unknown parameters of the model are estimated using a Bayesian approach. We apply a MCMC algorithm for a spatially homogeneous plot, study some statistical properties of associated estimators and present some numerical studies. Let us note in addition that even if our model is well adapted for a spatially homogeneous plot it can be extended for an orchard composed of several plots. Clearly in this case the implementation is more difficult and the time execution more expensive, which is beyond the scope of this paper.

The paper is organized as follows. The next section describes briefly the experiment and the data which will serve us to define the model and illustrate the estimation method. In sections 3 and 4 we introduce some notations and describe the model. Section 5 discusses simulation-based Bayesian inference and section 6 presents some numerical studies. Proofs and some technical details in relation with MCMC algorithm are postponed to the Appendix.

¹This work was supported by MODEMAVE research project from the Région Pays de la Loire.

2 Data description

The data provides from an experiment conducted in a research center of Angers² from 2004 to 2008 (Caffier et al. (2010), Didelot et al. (2010)). The objective of this experiment was to understand the spread of disease on resistant apple cultivar, Ariane, after overcoming of the resistance gene (Vf). First scab symptoms were observed in 2004 in this orchard. The orchard is composed of 16 plots separated by hedges that permits us to consider them to be spatially independent. It was partially observed 4 or 5 times a year. Each plot was composed of 78 trees regularly planted in 6 rows of 13 trees. In this paper we will concentrate our attention on three plots not treated against scab (A1, A2, A3) and composed of Ariane cultivar. More precisely we are interested in the spread of the inoculum over space and time. The observed variable will be called *disease score* and corresponds for each tree to the "visual estimation" of the proportion of scabbed leaves, a method adapted from Lateur et al (1998). Considering an increasing scale $0 < a_1 < a_2 < \dots < a_9 = 1$ an expert gives the score $j - 1$ to a tree when he estimates that the proportion of scabbed leaves belongs to the interval $[a_{j-1}, a_j]$. From a statistical point of view the observations are ordered categorical data. More precisely we will focus our attention to the study of the spread of the infection of a pure plot during a year. We present in figure 1 the disease score of the plot A2 at observation days (60, 66, 101, 138 and 185) during the year 2008. The model proposed in this paper considers that the plot is spatially homogeneous.

3 Notations

Let us consider a plot of trees of size $n = 78$ indexed by $i \in \mathcal{I} = \{1, \dots, n\}$ with a geographical structure which will permits us to define a neighborhood set. To each tree i we associate the disease score $Y_{i,t}$ at instant t . Let \mathbf{Y} be the *score process* $\mathbf{Y} = (Y_{i,t})_{i \in \mathcal{I}, t \in [0, EP]}$ taking values on $\{0, \dots, 8\}$ where EP represents the end of parasitic phase. This process is null at instant $\tau_0 = 0$ that is $Y_{i,0} \equiv 0 \forall i \in \mathcal{I}$ and observed at instants τ_1, \dots, τ_5 :

$$0 = \tau_0 < \tau_1 < \tau_2 < EA < \tau_3 < \tau_4 < \tau_5 = EP$$

where EA denotes the end of the infection providing from the ground (ascospore emission). The observed process at points τ_0, \dots, τ_5 will be denoted \mathbf{Y}^{obs} .

For each $i \in \mathcal{I}$, let us denote T_i^j the transition time that is the instant when the tree i changes the score from $j - 1$ to j and J_i the latest disease score reached by the tree i . So $Y_{i,EP} = J_i$, and:

$$\mathbf{T} = \{\mathbf{T}_i : i \in \mathcal{I}\} = \left\{ T_i^j, j = 1 \dots J_i : i \in \mathcal{I} \right\}. \quad (1)$$

\mathbf{T} is the set of unobserved transition instants which uniquely determines the score process known also as latent variables. For each $i \in \mathcal{I}$, we can then define the Poisson point process, with intensity given by (Daley and Vere-Jones, 2003):

$$\lambda_i(t) = E[dY_{i,t} \mid \mathbf{Y}_s : s < t] / dt. \quad (2)$$

Let $\partial(0)$ be some neighborhood set of a the origin point 0. Neglecting boundary effects for some of the trees of the plot, the neighbors of a tree i are $\partial(i) = i + \partial(0)$. We suppose that $\partial(0)$ can be written as the disjoint union of neighbors of type v , $\partial(0) = \bigcup_{v=1}^{N_v} \partial_v(0)$ which defines the function $\mathbf{v} : \partial(0) \rightarrow \{1, \dots, n_V\}$ such that $\mathbf{v}(\ell) = v$ for each $\ell \in \partial_v(0)$.

We consider the following observed variables (covariates):

- The PAD (potential ascospore dose) is a scalar giving an estimation of inoculum available on the ground which starts up the infection. The value of PAD is estimated for each plot and is considered to be homogeneous within each plot. It represents the stock of pathogen for primary infections.
- Every day d is characterized by one of $n_{\mathbf{R}}$ levels of risk, denoted $0, \dots, n_{\mathbf{R}} - 1$. The risk is obtained by Pulsowin 3.1 software³ which is based on Mill's curves (Mills and Laplante, 1951 and Olivier, 1986) and distinguish 5 levels of risks depending on the temperature and the length of leaf humectation. It is supposed to remain constant during a day. We will denote \mathbf{R} the vector of length EP which d -th component corresponds to the *risk* of the infection of the day d .

²INRA, UMR PAVE, 42 rue Georges Morel, F-49071 Beaucouz

³Pulsonic, Orsay, France

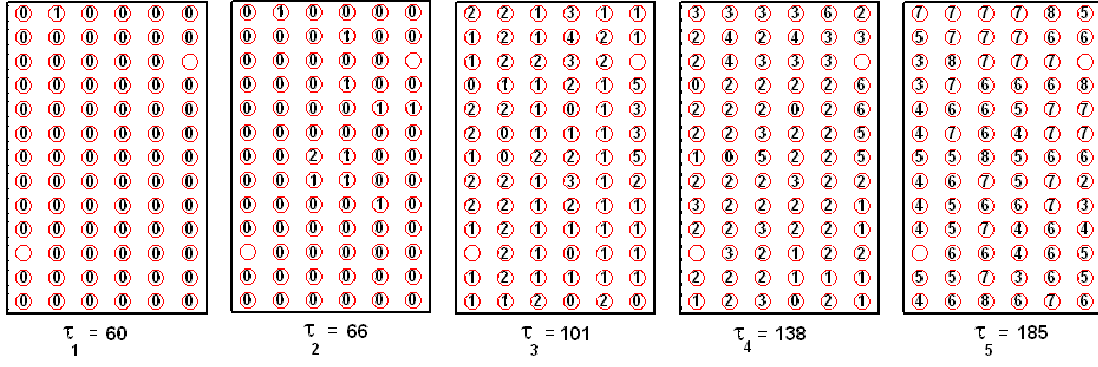


Figure 1: Disease scores for the plot A2 at observed instants τ_1, \dots, τ_5 .

4 The model

Let us denote $\mathbf{Z} = (\mathbf{R}, PAD, \mathbf{a})$ where $\mathbf{a} = (a_1, \dots, a_9)$. We suppose that the process \mathbf{Y}^{obs} is a sample of a Poisson point process with intensity $\lambda_i(t, \boldsymbol{\theta}; \mathbf{Z})$ for the tree i at instant t . This intensity is obtained as the product of two terms:

$$\lambda_i(t, \boldsymbol{\theta}; \mathbf{Z}) = r(\alpha_{\mathbf{R}_{E(t)}}) \times \left[C(\gamma) \times \rho(t) + g_i(t; \mathbf{T}_i, \mathbf{a}) + \sum_{\ell \in \partial(0)} g_{i+\ell}(t; \mathbf{T}_{i+\ell}, \mathbf{a}) \times K(\ell, \eta_{\mathbf{v}(\ell)}) \right] \quad (3)$$

where $E(\cdot)$ denotes the integer part function.

The first term $r(\alpha_{\mathbf{R}_{E(t)}})$ describes climatic risk of infection for the day $E(t)$, since as previously mentioned the risk remains constant during a day. The second term describes the dispersion process of *primary* and *secondary infection*, which represents the pathogen produced by lesions on trees. $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \gamma, \boldsymbol{\eta})$ are the unknown parameters providing the contribution of different biological phenomena. The vector parameter $\boldsymbol{\alpha}$ corresponds to weather conditions, the scalar parameter γ corresponds to the primary infection and the vector parameter $\boldsymbol{\eta}$ corresponds to the secondary infection. We will present now the functions r, C, ρ, g_i and K of the expression (3). This intensity can be seen as the superposition of two point processes ([14]), where the first one represents the risk of infection by primary inoculum initiating the epidemics. Then the second process appears when the first tree is infected.

$r(\cdot)$ is a positive and increasing function and $\boldsymbol{\alpha}$ a vector parameter of $n_{\mathbf{R}}$ components $\alpha_0, \dots, \alpha_{n_{\mathbf{R}}-1}$ satisfying the condition $\alpha_0 < \dots < \alpha_{n_{\mathbf{R}}-1}$.

The primary infection is given by the product $C(\gamma) \times \rho(t)$. $C(\gamma)$ is a positive and bijective function representing the infection providing from the ground. $\rho(t)$ describes the quantity of available infection in the ground at instant t all over the plot. It means that we suppose that this quantity is spatially uniformly distributed. Clearly $\rho(t)$ is a positive function defined on $[0, EA]$ since this phenomena is possible only during this period.

The crucial question of this paper concerns the spread of secondary infection that is the spread of infection between trees. In order to propose a solution to this question we will firstly introduce the function g_i which estimates the proportion of scabbed leaves of the tree i at each instant $t \in [0, EP]$. Secondly we will introduce the term K for modeling the spread of infection between trees.

Recall that the proportion of infected leaves for each tree is partially known only at instants τ_0, \dots, τ_5 : for instance, if $Y_{i, \tau_1} = 1$ then the value of the proportion of infected leaves at instant $t = \tau_1$ belongs to the interval $[a_1, a_2)$. The function g_i is such that $g_i(T_i^j; \mathbf{T}_i, \mathbf{a}) = a_j$ for $j = 1, \dots, J_i$. Let us define $g_i(T_i^{J_i+1}; \mathbf{T}_i, \mathbf{a}) = (a_{J_i} + a_{J_i+1})/2$. Supposing, as for the logistic growth models (Madden, Hughes, Van den Bosch, 2007 or Verhulst, 1845), that the proportion of infected leaves is proportional to the amount of leaves affected and to the number of leaves that can still be affected, a direct calculation leads to the following

result:

$$g_i(t; \mathbf{T}_i, \mathbf{a}) = \begin{cases} 0 & t < T_i^1 \\ \text{logit}^{-1}\left\{\frac{\tilde{a}_{j+1}-\tilde{a}_j}{T_i^{j+1}-T_i^j}(t-T_i^j) + \tilde{a}_j\right\} & T_i^j \leq t \leq T_i^{j+1}, j = 1 \dots J_i \end{cases} \quad (4)$$

where $\text{logit}(x) = \log[x/(1-x)]$ for $x \in (0, 1)$, $\tilde{a}_j = \text{logit}(a_j)$ for $j = 1 \dots J_i$, $\tilde{a}_{J_i+1} = \text{logit}[(a_{J_i} + a_{J_i+1})/2]$ and $T_i^{J_i+1} = EP$.

The function g_i is zero in the interval $[0, T_i^1]$ and strictly increasing on $[T_i^1, EP]$ for each tree i such that $J_i > 0$. Furthermore it satisfies differential equation $\frac{\partial g(t; \mathbf{T}_i, \mathbf{a})}{\partial t} = \beta g(t; \mathbf{T}_i, \mathbf{a}) (1 - g(t; \mathbf{T}_i, \mathbf{a}))$ for $t \in [T_i^1, EP]$. Note finally that alternative values are possible to define $g_i(T_i^{J_i+1}; \mathbf{T}_i, \mathbf{a})$. The figure (2) represents the function g_i of a tree i with simulated instants t_i^j .

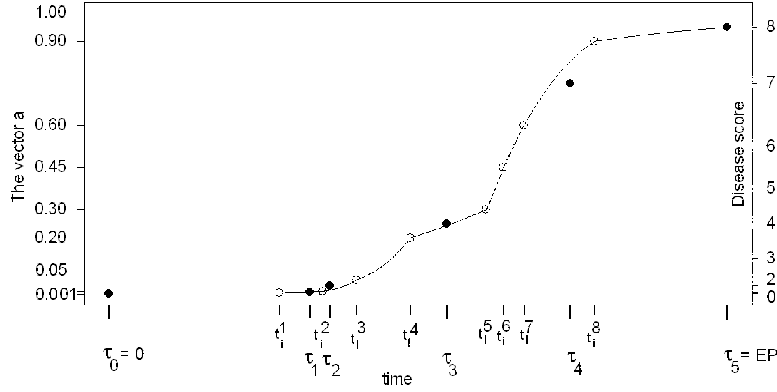


Figure 2: Observations (filled points) and an estimation of the proportion of infected leaves represented by the function g_i (solid line) such that $g_i(t_i^j; \mathbf{T}_i, \mathbf{a}) = a_j$ for $j = 1, \dots, J_i$ (empty points).

Let ℓ be a neighbor (of 0) of type $v = \mathbf{v}(\ell)$. The function $K(\ell, \eta)$ describes the dispersion process and may be any suitable function with $\eta = \eta_v$ as unknown parameter for characterizing the type v of the neighborhood set. We denote $\boldsymbol{\eta}$ the vector $(\eta_v, v = 1, \dots, n_V)$ with n_V components of unknown parameters. Clearly the choice of the distance and neighborhood set depends on the application.

The next paragraph states the conditions under which the intensity parameter of the underlying Poisson process is uniquely determined.

4.1 Identifiability conditions

The model 3 is well defined if $\lambda_i(t, \boldsymbol{\theta}; \mathbf{Z})$ is an injective function of $\boldsymbol{\theta}$ that is if $\boldsymbol{\theta}, \boldsymbol{\theta}'$ are two points of $\boldsymbol{\theta}$ then :

$$\{(\lambda_i(t, \boldsymbol{\theta}; \mathbf{Z}) = \lambda_i(t, \boldsymbol{\theta}'; \mathbf{Z})) \forall i \in \mathcal{I}, \forall t \in [0, EP]\} \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}' \quad (5)$$

We prove the identifiability property in appendix under the following assumptions.

Condition 1 *At least a tree i_1 is contaminated during $[0, EA]$*

Condition 2 *For each $v = 1, \dots, n_V$ there exists a tree ℓ having i_1 as a neighbor of type v . In particular when the neighborhood set is symmetrical this condition is equivalent to: the tree i_1 has neighbors of each type v*

Condition 3 *The function ρ is strictly positive in $[0, EA]$ and not proportional to any of the growth functions $\{g_i : i \in \mathcal{I}\}$ at any interval δ of $[0, EA]$ with positive length*

Condition 4 *For every $\ell \in \partial(0)$ the dispersal function $K(\ell, \eta)$ is an injective function with respect to η*

Condition 5 *The function $C(\gamma)$ is a bijective and strictly positive function with respect to γ*

The condition 1 is necessary since the infection can start up only during the interval $(0, EA]$. The condition 2 can be written:

$$\forall v \in \{1, \dots, n_V\} \exists \ell \in \mathcal{I} : i_1 \in \partial_v(\ell) \quad (6)$$

and in the symmetrical case:

$$\forall v \in \{1, \dots, n_V\} \exists \ell \in \partial_v(i_1)$$

The third condition is a slight restriction in the definition of the function ρ . The condition 4 is natural since K is another way of the characterization of the type of a neighbor.

Let now give the identification property which is proved in appendix.

Lemma 6 *The model (3) is well defined under conditions 1 to 5.*

5 MCMC estimation of the parameters

The estimation of the parameters will be based on the maximisation of

$$P_{\boldsymbol{\theta}}(\mathbf{Y}^{\text{obs}} \mid \mathbf{Z}) = \int_{\mathbf{T} \in \mathcal{T}} dP_{\boldsymbol{\theta}}(\mathbf{Y}^{\text{obs}}, \mathbf{T} \mid \mathbf{Z}) = \int_{\mathcal{T}} P(\mathbf{Y}^{\text{obs}} \mid \mathbf{T}, \mathbf{Z}) P_{\boldsymbol{\theta}}(\mathbf{T} \mid \mathbf{Z}) dP(\mathbf{T}). \quad (7)$$

As this expression is intractable, we adopt a Bayesian approach, considering $\boldsymbol{\theta}$ as a random variable (Robert et al., 2004) by applying a MCMC based procedure using a Metropolis-Hastings algorithm where transition times \mathbf{T} are considered as latent variables. This approach has been already used in plant pathology domain when \mathbf{Y} is a binary valued process and when transition instants are not observed (see for instance Gibson, 1997; Gamerman and Lopes, 2006; Rasmussen et al., 2007; Soubeyrand et al., 2009 among others). We generalize this approach in the case where \mathbf{Y} is an ordinal categorical variable.

Considering $\boldsymbol{\theta}$ as a random variable the posterior distribution of $\boldsymbol{\theta}, \mathbf{T}$, up to a normalizing constant, can be written as:

$$P(\mathbf{T}, \boldsymbol{\theta} \mid \mathbf{Y}^{\text{obs}}, \mathbf{Z}) \propto P(\mathbf{Y}^{\text{obs}} \mid \mathbf{T}, \boldsymbol{\theta}, \mathbf{Z}) P(\mathbf{T} \mid \boldsymbol{\theta}, \mathbf{Z}) \pi(\boldsymbol{\theta}).$$

Recall that \mathbf{T}_i is the vector of transition instants (T_i^j) at which the tree i changes the score from $j-1$ to j for $j = 1, \dots, J_i$. Making an appropriate choice of \mathbf{T}_i for $i \in \mathcal{I}$ leads to $P(\mathbf{Y}^{\text{obs}} \mid \mathbf{T}, \boldsymbol{\theta}, \mathbf{Z}) = 1$ which simplifies the algorithm. The expression of $P(\mathbf{T} \mid \boldsymbol{\theta}, \mathbf{Z})$ is given by the following lemma and the prior distribution $\pi(\boldsymbol{\theta})$ will be described in the next paragraph.

Lemma 7 *Let $\mathbf{t} = \{t_i^j : \forall i, j\}$ be a sample of transition events. Then:*

$$P(\mathbf{T} = \mathbf{t} \mid \boldsymbol{\theta}, \mathbf{Z}) = P\left(T_i^j = t_i^j : \forall i, j \mid \boldsymbol{\theta}, \mathbf{Z}\right) = \prod_{i \in \mathcal{I}} \left[e^{-\Lambda_i(0, EP)} \prod_{j=1}^{J_i-1} \lambda_i(t_i^j) \right] \quad (8)$$

where the simplified notation $\forall i, j$ is used instead of $i \in \mathcal{I}, j = 1, \dots, J_i$ and for $u < v$ and $i \in \mathcal{I}$:

$$\Lambda_i(u, v) = \int_u^v \lambda_i(t) dt. \quad (9)$$

Note that in the above lemma, proved in the appendix, we have abused notation $P(T_i^j = t_i^j : \forall i, j \mid \boldsymbol{\theta}, \mathbf{Z})$ for the density of the process under discussion. In the following we will use the simplified notation $P(\mathbf{t} \mid \boldsymbol{\theta}, \mathbf{Z})$ and:

$$P(\mathbf{t}_A \mid \boldsymbol{\theta}, \mathbf{Z}) = \prod_{i \in A} \left[e^{-\Lambda_i(0, EP)} \cdot \prod_{j=1}^{J_i} \lambda_i(t_i^j) \right]$$

for any subset A of \mathcal{I} .

5.1 Sequential rules for the updating of parameters and transition events

In this paragraph we outline the MCMC algorithm carried out in two steps. In a first time we update transition times by considering in order all the trees. In a second time we update each component α , γ , η of the parameter θ based on the decomposition:

$$P(\theta, \mathbf{T} \mid \mathbf{Y}^{obs}, \mathbf{Z}) \propto P(\mathbf{T} \mid \theta, \mathbf{Z}) \pi_{\alpha}(\cdot) \pi_{\gamma}(\cdot) \pi_{\eta}(\cdot).$$

Let us give a brief description of the different steps of the algorithm.

We will denote $\pi_{\alpha}(\cdot)$, $\pi_{\gamma}(\cdot)$, $\pi_{\eta}(\cdot)$ the prior distributions and $q_{\alpha}(\cdot|\cdot)$, $q_{\gamma}(\cdot|\cdot)$, $q_{\eta}(\cdot|\cdot)$ the proposal distributions for the corresponding parameters. Details about these distributions are postponed in the appendix.

For each $i \in \mathcal{I}$, let \mathbf{t}_i^c be the current value of transition times and \mathbf{t}_i^* the candidate value furnished by proposal distribution $q_{\mathbf{t}}(\cdot|\cdot)$ described in the appendix. Similarly let $\theta^c = (\alpha^c, \gamma^c, \eta^c)$ be the current value of the parameter θ and θ^* the candidate value obtained by the associated proposal distributions.

For each $i \in \mathcal{I}$ we replace \mathbf{t}_i^c by \mathbf{t}_i^* with probability:

$$\min \left\{ 1, \frac{P(\mathbf{t}_i^*, \mathbf{t}_{\mathcal{I} \setminus i}^c \mid \theta^c, \mathbf{Z}) q(\mathbf{t}_i^c \mid \mathbf{t}_i^*)}{P(\mathbf{t}_i^c, \mathbf{t}_{\mathcal{I} \setminus i}^c \mid \theta^c, \mathbf{Z}) q(\mathbf{t}_i^* \mid \mathbf{t}_i^c)} \right\} = \min \left\{ 1, \frac{P(\mathbf{t}_i^*, \mathbf{t}_{\partial(i)}^c \mid \theta^c, \mathbf{Z}) q(\mathbf{t}_i^c \mid \mathbf{t}_i^*)}{P(\mathbf{t}_i^c, \mathbf{t}_{\partial(i)}^c \mid \theta^c, \mathbf{Z}) q(\mathbf{t}_i^* \mid \mathbf{t}_i^c)} \right\} \quad (10)$$

Actually for both transition times $(\mathbf{t}_i^*, \mathbf{t}_{\mathcal{I} \setminus i}^c)$ and $(\mathbf{t}_i^c, \mathbf{t}_{\mathcal{I} \setminus i}^c)$ the associated intensities $\lambda_{\ell}(t_{\ell}^j)$ are identical for $\ell \in \mathcal{I} \setminus \mathcal{N}_i$ where $\mathcal{N}(i) = \{i\} \cup \partial(i)$. The identity (10) follows since $P(\mathbf{t}_i^*, \mathbf{t}_{\mathcal{I} \setminus i}^c \mid \theta^c, \mathbf{Z}) = P(\mathbf{t}_i^*, \mathbf{t}_{\partial(i)}^c \mid \theta^c, \mathbf{Z}) \cdot P(\mathbf{t}_{\mathcal{I} \setminus \mathcal{N}_i}^c \mid \theta^c, \mathbf{Z})$ and $P(\mathbf{t}_i^c, \mathbf{t}_{\mathcal{I} \setminus i}^c \mid \theta^c, \mathbf{Z}) = P(\mathbf{t}_i^c, \mathbf{t}_{\partial(i)}^c \mid \theta^c, \mathbf{Z}) \cdot P(\mathbf{t}_{\mathcal{I} \setminus \mathcal{N}_i}^c \mid \theta^c, \mathbf{Z})$.

The parameters are updated component by component in the order appearing in θ . The two first components α and γ are updated in a similar way. For instance the current value α^c is replaced by α^* with probability:

$$\min \left\{ 1, \frac{P(\mathbf{t}^c \mid \alpha^*, \gamma^c, \eta^c, \mathbf{Z}) \pi_{\alpha}(\alpha^*) q(\alpha^c \mid \alpha^*)}{P(\mathbf{t}^c \mid \alpha^c, \gamma^c, \eta^c, \mathbf{Z}) \pi_{\alpha}(\alpha^c) q(\alpha^* \mid \alpha^c)} \right\} \quad (11)$$

As there are no constraints on the components of the vector η we update sequentially all these components. Using the simplifying notation η_v^a, η_v^b for $\eta_1^a, \dots, \eta_{v-1}^a, \eta_v^b, \eta_{v+1}^a, \dots, \eta_{n_v}^a$ the candidate value η_v^* replaces the current value η_v^c with probability:

$$\min \left\{ 1, \frac{P(\mathbf{t} \mid \alpha^c, \gamma^c, \eta_v^c, \eta_v^*, \mathbf{Z}) \pi_{\eta}(\eta_v^c, \eta_v^*) q(\eta_v^c \mid \eta_v^c, \eta_v^*)}{P(\mathbf{t} \mid \alpha^c, \gamma^c, \eta_v^c, \eta_v^c, \mathbf{Z}) \pi_{\eta}(\eta_v^c, \eta_v^c) q(\eta_v^c, \eta_v^* \mid \eta_v^c)} \right\} \quad (12)$$

6 Numerical results

We will now present some numerical studies in relation with the data introduced in the 2nd section carried out with R software⁴. We define r as the logit function and C as the exponential of $\gamma * PAD$. Following works of Aylor (1998, 1999), we suppose that the quantity of ground infection $\rho(t)$ is dispersed in time following the density of a truncated Gaussian variable (Johnson et al., 1995) defined on $[0, EA]$ centered in $\frac{EA}{2}$ and standard error σ_{ρ} . The parameter σ_{ρ} will serve as a tuning parameter in MCMC simulations. The dispersal function K is given by:

$$K(\ell, \eta_{\mathbf{v}(\ell)}) = \exp[-\eta_{\mathbf{v}(\ell)} \cdot d(0, \ell)]$$

where $d(i, \ell)$ is the Euclidean distance between trees i and ℓ . We apply the model (3) with initial neighborhood $\partial(0) = \{1, 2, 3, 4\}$ as illustrated in figure 3. The type of the neighbor ℓ is $\mathbf{v}(\ell) = \ell$ and so $\eta = (\eta_1, \eta_2, \eta_3, \eta_4)$. We run a MCMC procedure with 50.000 iterations and burn-in-length of 8.000. The choice of neighbors were determined by applying two criteria : the convergence and stability of posterior distribution of all the components of θ . Neighbors whose dispersal parameter do not converge are removed from the model, meaning that they can be omitted. The final neighborhood set for all the plots is composed of right (1) and upper (2) neighbors that is $\eta = (\eta_1, \eta_2)$. As expected for the final model, the simulations confirm the independence of posterior distributions from the initial values $\theta_o = (\alpha_o, \gamma_o, \eta_o)$ of θ and \mathbf{T}_o of \mathbf{T} . We present here numerical results with $\alpha_o = (0.01, 0.02, 0.03, 0.04, 0.05)$, $\gamma_o = -6$ and $\eta_o = (10, 10)$. Even though

⁴cran.r-project.org

our main interest concerns the mean values of each scalar parameter, we firstly pay attention to assess the statistical property of simulated chains by presenting some detailed results for the plot A2 and then give some summary results for the three plots.

The results concerning the parameter η for the plot A2 are presented in figures 3 and 4. Clearly the convergence properties are better satisfied for the model with neighborhood set defined with neighbors 1 and 2. For this model the density estimation for the two parameters η_1 and η_2 seems to be approximately Gaussian as the theory tells us (see Th. 2.1, Geyer, 1992). In addition we calculate three initial sequence estimators of the standard error of MCMC estimators, called positive, monotone and convex with $lag = 453$ as described in Geyer (Th. 3.2, 1992) and present them in Table 1. The proximity of three estimations is another satisfactory property which confirms the choice of final neighborhood composed of two sites.

We use the "positive" initial sequence estimators in the presentation of numerical results for the plot A1, A2, A3 as suggested in general case (see for example Gelman and Rubin, 1992). A summary of the parameter estimations for plots A1, A2 and A3 in the case of two neighbors is given in Table 2. It gives the estimations (mean of simulated chains) and the functions $K(1, \eta_1)$, $K(2, \eta_2)$ corresponding to θ and based on 95% confidence intervals (CI).

To conclude, for all plots the most influential neighbors are the right and upper (row) ones which match up with the main wind direction during the saison. Their influence depends on local position of the plot in the orchard and the level of PAD.

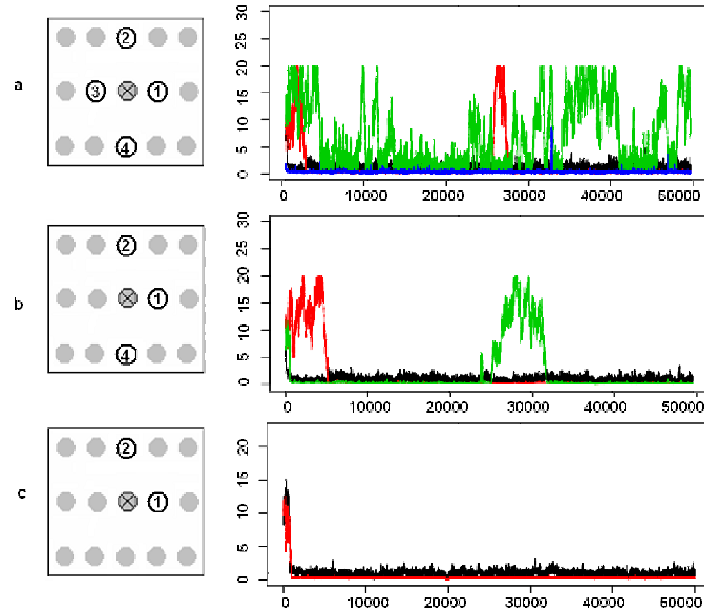


Figure 3: Left: Neighborhood configurations with respectively 4, 3 and 2 sites. Right: Simulated chains corresponding to parameters η_1 , η_2 for the plot A2 with different neighborhood set: (a) $\sigma(0) = \{1, 2, 3, 4\}$, (b) $\sigma(0) = \{1, 2, 4\}$, (c) $\sigma(0) = \{1, 2\}$.

Parameter	Mean estimation	Initial sequence estimators for SD		
		Positive	Monotone	Convexe
η_1	0.9994	0.0235	0.0232	0.0226
η_2	0.2882	0.0050	0.0049	0.0047

Table 1: Estimations of parameters η_1 , η_2 (mean of simulated chains) and different initial sequence estimates of the standard error (SD) of the mean estimator for the plot A2.

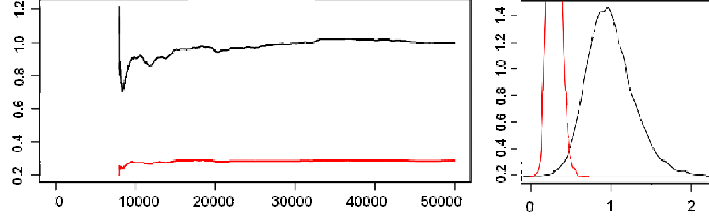


Figure 4: Mean and histogram of simulated chain corresponding to η_1 and η_1 for the plot A2 (burn-in-length of 8000)

Plot	α_0	α_1	α_2	α_3	α_4	γ	η_1	η_2	$K(1, \eta_1)$	$K(2, \eta_2)$
A1	-1.5180	-1.4148	-1.3374	-1.2304	-1.1036	0.3366	0.9269	0.3257	0.3139	0.2717
A2	-1.5066	-1.4249	-1.3780	-1.2957	-1.2004	0.5311	0.9994	0.2882	0.2867	0.3158
A3	-2.6022	-2.5393	-2.4867	-2.4110	-2.3233	2.6413	0.4750	0.0605	0.5523	0.7850

Table 2: Estimation of parameters and some of their corresponding functions for the three plots

Conclusion

We present a spatial-temporal point process for modeling the spread of apple scab within an orchard composed of several plots. The process is indexed by the sites of a fixed spatial lattice and continuous in time. We estimate the parameters by using a Bayesian approach and implementing a MCMC algorithm. Based on observations of the orchard at sparsely distributed times we present some numerical results that permits us to determine the most infectious neighbors and quantify the probability of infection from one tree to a neighbor. In order to increase the power of statistical inference our approach permits also to identify similar plots that can be treated together in analyzing the data.

A Appendix

A.1 Proof of lemma 6

Omitting \mathbf{Z} in (3) let $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ be such that:

$$\lambda_i(t, \boldsymbol{\theta}) = \lambda_i(t, \boldsymbol{\theta}') \quad \forall i \in \mathcal{I}, \forall t \in [0, EP].$$

Let i_1 be the first infected tree satisfying the condition C1 at instant $t_{i_1}^1$, $R = \mathbf{R}_{E(t_{i_1}^1)}$ and $\delta_{i_1} = [t_{i_1}^1, \min(E(t_{i_1}^1) + 1, t_{i_2}^{j_2})]$. From (3) and as $g_\ell(t) = 0 \quad \forall t \in \delta_{i_1}$ and $\ell \in \partial(i_1)$ it follows that:

$$\rho(t) [r(\alpha_R)C(\gamma) - r(\alpha'_R)C(\gamma')] + g_{i_1}(t) [r(\alpha_R) - r(\alpha'_R)] = 0 \quad \forall t \in \delta_{i_1}. \quad (13)$$

If $\alpha_R \neq \alpha'_R$ then $r(\alpha_R) \neq r(\alpha'_R)$ and so:

$$g_{i_1}(t) = \rho(t) \frac{[r(\alpha_R)C(\gamma) - r(\alpha'_R)C(\gamma')]}{[r(\alpha_R) - r(\alpha'_R)]} \quad \forall t \in \delta_{i_1}$$

which is in contradiction with condition 3 since δ_{i_1} is an intervalle with positive length.

Equation (13) when $\alpha_R = \alpha'_R$ implies:

$$\rho(t) [C(\gamma) - C(\gamma')] = 0 \quad \forall t \in \delta_{i_1}.$$

As ρ is strictly positive in $]0, EA]$ and $C(\gamma)$ is a bijective function one obtains $C(\gamma) = C(\gamma')$ and so $\gamma = \gamma'$.

Let now consider a neighbor ℓ_0 of i_1 of type v_o . The tree i_1 being the unique infected tree since $g_{\ell_0}(t) = 0$ for $t \in \delta_{i_1}$ one gets:

$$\sum_{\ell \in \partial(\ell_0)} g_\ell(t) \left[K(\ell - \ell_0, \eta_{\mathbf{v}(\ell - \ell_0)}) - K(\ell - \ell_0, \eta'_{\mathbf{v}(\ell - \ell_0)}) \right] = 0 \quad \forall t \in \delta_{i_1}$$

that is, as i_1 is a neighbor of type v_o of ℓ_0

$$g_{i_1}(t) \left[K(i_1 - \ell_0, \eta_{v_o}) - K(i_1 - \ell_0, \eta'_{v_o}) \right] = 0 \quad \forall t \in \delta_{i_1}$$

which implies $\eta_{v_o} = \eta'_{v_o}$ since K is an injective function from condition 4. As from condition 2 there exists at least one neighbor of each type it follows that $\eta_v = \eta'_v$ for all $v \in \{1, \dots, n_V\}$.

From condition ?? let fixe a day d with risk R_o and an instant t_d of this day. Since $\gamma = \gamma'$ and $\eta_v = \eta'_v$ if $\lambda_{i_1}(t_d, \boldsymbol{\theta}) = \lambda_{i_1}(t_d, \boldsymbol{\theta}')$ then:

$$(r(\alpha_{R_o}) - r(\alpha'_{R_o})) \left\{ C(\gamma)\rho(t_d) + g_{i_1}(t_d) + \sum_{\ell \in \partial(i_1)} g_\ell(t_d)K(\ell, \eta_{\mathbf{v}(\ell)}) \right\} = 0$$

But $\gamma\rho(t_d) + g_{i_1}(t_d) + \sum_{\ell \in \partial(i_1)} g_\ell(t_d)K(\ell, \eta_{\mathbf{v}(\ell)})$ is strictly positive. Indeed if $t_d < EA$ then $C(\gamma)\rho(t_d) > 0$ and if $t_d \geq EA \geq t_{i_1}^1$ then $g_{i_1}(t_d) > 0$. Consequently $r(\alpha_{R_o}) = r(\alpha'_{R_o})$ and $\alpha_{R_o} = \alpha'_{R_o}$ which completes the proof.

A.2 Proof of Lemma 7

We will omit $\boldsymbol{\theta}$ in this paragraph. Recall that $\forall i, j$ means $i \in \mathcal{I}, j = 1, \dots, J_i$. Let us denote

$$\mathbf{P}_{\mathbf{T}} = P \left(T_i^j = t_i^j : \forall i, j \right)$$

and $N_{\mathbf{J}} = \sum_{i=1, n} J_i$ the total number of transition instants. Let us consider the ordered sequence of transitions instants $t_{(1)} < \dots < t_{(N_{\mathbf{J}})}$ such that $t_{(k)} = T_{i_k}^{j_k}$ for some i_k, j_k . Writing $t_{(0)} = 0$ and $t_{(N_{\mathbf{J}}+1)} = EP$ and using the Markovian property we get:

$$\begin{aligned} \mathbf{P}_{\mathbf{T}} &= P \left(T_i^j = t_i^j : \forall i, j \right) \\ &= P \left(T_{i_k}^{j_k} = t_{(k)} : k = 1, \dots, N_{\mathbf{J}} \right) \\ &= \prod_{k=1}^{N_{\mathbf{J}}+1} P \left(T_{i_k}^{j_k} = t_{(k)} \middle| T_{i_l}^{j_l} = t_{(l)} : l < k \right) \\ &= \prod_{k=1}^{N_{\mathbf{J}}+1} P \left(T_{i_k}^{j_k} = t_{(k)} \middle| T_{i_{k-1}}^{j_{k-1}} = t_{(k-1)} \right). \end{aligned} \tag{14}$$

Let fixe some k in $\{1, \dots, N_{\mathbf{J}}\}$. The event $A_k : "T_{i_k}^{j_k} = t_{(k)} | T_{i_{k-1}}^{j_{k-1}} = t_{(k-1)}"$ means "during the interval $(t_{(k-1)}, t_{(k)})$ only one event takes place : the score level of the tree i_k increase by one at instant $t_{(k)}$; no other changes take place during this interval". It follows that (see for exemple Daley et Jones, Vol; 1, pg. 23):

$$P \left(T_{i_k}^{j_k} = t_{(k)} \middle| T_{i_{k-1}}^{j_{k-1}} = t_{(k-1)} \right) = \lambda_{i_k}(t_{(k)}) \cdot \prod_{i \in \mathcal{I}} e^{-\Lambda_i(t_{(k-1)}, t_{(k)})} \tag{15}$$

since the successive events A_k are independent for the different trees where $\Lambda_i(t_{(k-1)}, t_{(k)}) = \int_{t_{(k-1)}}^{t_{(k)}} \lambda_i(t) dt$. In particular for the event $A_{N_{\mathbf{J}}+1}$ that is "No changes take place between $t_{(N_{\mathbf{J}})}$ and $t_{(N_{\mathbf{J}}+1)} = EP$ " we have:

$$P(A_{N_{\mathbf{J}}+1}) = \prod_{i \in \mathcal{I}} e^{-\Lambda_i(t_{(N_{\mathbf{J}})}, t_{(N_{\mathbf{J}}+1)})} \tag{16}$$

The result is a consequence of $\Lambda_i(0, EP) = \sum_{k=1}^{N_J+1} \Lambda_i(t_{(k-1)}, t_{(k)})$ and equations 14, 15 et 16 since:

$$\begin{aligned} \mathbf{P}_{\mathbf{T}} &= \prod_{k=1}^{N_J+1} P\left(T_{i_k}^{j_k} = t_{(k)} \middle| T_{i_{k-1}}^{j_{k-1}} = t_{(k-1)}\right) \\ &= \prod_{k=1}^{N_J+1} \left[\lambda_{i_k, t_{(k)}} \prod_{i \in \mathcal{I}} e^{-\Lambda_i(t_{(k-1)}, t_{(k)})} \right] \\ &= \prod_{i \in \mathcal{I}} \left[\prod_{j=1}^{J_i} \lambda_i(t_i^j) \cdot \prod_{k=1}^{N_J+1} e^{-\Lambda_i(t_{(k-1)}, t_{(k)})} \right] \\ &= \prod_{i \in \mathcal{I}} \left[\prod_{j=1}^{J_i} \lambda_i(t_i^j) \cdot e^{-\Lambda_i(0, EP)} \right]. \end{aligned}$$

A.3 Priors and proposals

As no information is available we consider uniform prior distribution for the components $\gamma, \boldsymbol{\eta}$ of the unknown parameter $\boldsymbol{\theta}$. Their support are respectively $[-10, 10]$ for γ and $[0, 20]$ for each component of $\boldsymbol{\eta}$. $\boldsymbol{\alpha}$ satisfy the property $\alpha_0 > \dots > \alpha_{n_{\mathbf{R}}-1}$. Similarly we consider uniform priors for the components of $\boldsymbol{\alpha}$ with supports that assure the order property. The support for α_0 is $[-5, 5]$. In a similar way we define uniform priors for the transition times which satisfy two conditions. They satisfy order property for each tree $i \in \mathcal{I}$, that is $t_i^j < t_i^{j+1}, \forall j$ and the number of transitions times between two consecutive observation instants τ_{k-1}, τ_k equals $Y_{i,k+1} - Y_{i,k}$.

The proposal distributions are based on truncated Gaussian laws (Johnson and Kotz, 1995) centered at the current value of the parameter. We denote $Z_{a,b}(\mu, \sigma)$ a truncated normal distribution centered in μ , standard error σ truncated at a and b and $\mathbf{z}_{a,b}(x; \mu, \sigma)$ his density at a point x .

The proposal distributions for γ and η_v for $v = 1, \dots, n_V$ are respectively $\mathbf{z}_{-10,10}(\gamma, \sigma_\gamma)$ and $\mathbf{z}_{-10,10}(\eta_j, \sigma_{\eta_j})$ for $v = 1, \dots, n_V$.

Let $\boldsymbol{\alpha}^* = (\alpha_0^*, \dots, \alpha_{n_{\mathbf{R}}-1}^*)$ be the candidate value where $\alpha_0^* \sim Z_{-5,5}(\alpha_0, \sigma_{\alpha_0})$ and $\alpha_j^* \sim Z_{\alpha_{j-1}^*, 5}(\alpha_j, \sigma_{\alpha_j})$. The proposal of $\boldsymbol{\alpha}$ is given by:

$$q(\boldsymbol{\alpha}^* | \boldsymbol{\alpha}) = \mathbf{z}_{-5,5}(\alpha_0^*; \alpha_0, \sigma_{\alpha_0}) \cdot \prod_{j=1}^{n_{\mathbf{R}}-1} \mathbf{z}_{\alpha_{j-1}^*, 5}(\alpha_j^*; \alpha_j, \sigma_{\alpha_j}) \quad (17)$$

Note that when $t_i^j \in [\tau_{k-1}, \tau_k)$ the candidate $t_i^{j,*}$ belongs also to the same interval $[\tau_{k-1}, \tau_k)$. For each tree $i \in \mathcal{I}$ such that that $J_i > 1$ the proposal distribution is given by:

$$q(\mathbf{t}_i^* | \mathbf{t}_i) = \mathbf{z}_{\max(0, \tau_{k_{i,1}-1}), \tau_{k_{i,1}}} (t_i^{1,*}; t_i^1, \sigma_{T^1}) \prod_{j=2}^{J_i-1} \mathbf{z}_{\max(t_i^{j-1,*}, \tau_{k_{i,j-1}}), \tau_{k_{i,j}}} (t_i^{j,*}; t_i^j, \sigma_{T^j}) \quad (18)$$

where $k_{i,j} = \min_k \{k : \tau_k \geq t_i^j\}$. When $J_i = 1$ the proposal is defined only by the first term.

Let us note that we have in our possession the tuning parameters, $\sigma_{\alpha_j}, j = 0, \dots, n_{\mathbf{R}} - 1, \sigma_\gamma, \sigma_{\eta_v}, v = 1, \dots, n_V, \sigma_{T^j}, j = 1 \dots 8$ to reach the required accepting rate for updated values for each of the parameters and transition times. In our simulations we chose values producing accepting rates approaching 0.23 (Roberts et al., 1997).

References

- [1] Aylor D.E. (1998) The aerobiology of apple scab. *Plant Disease*, 82, 8, 838-849.
- [2] Aylor D.E. (1999) Biophysical scaling and the passive dispersal of fungus spores: relationship to integrated pest management. *Agricultural and forest meteorology*, 97, 275-292.

- [3] Chadoeuf J., Nandris D., Geiger J.P., Nicole M., Pierrat J.C. (1992) Modélisation spatio-temporelle d'une épidémie par un processus de Gibbs: Estimation et Tests. *Biometrics*, 48, 1165-1175.
- [4] Cressie N., Wikle C.K. (2011) *Statistics for spatio-temporal data*. Wiley.
- [5] Daley D.J., Vere-Jones D. (2003) *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods, Second Edition*. Springer-Verlag
- [6] Caffier V, Didelot F, Orain G, Lemarquand A, Parisi L (2010). Efficiency of association of scab control methods on resistance durability of apple: the case study of cultivar Ariane. *IOBC wprs Bulletin*, 54, 327-330.
- [7] Didelot F, Caffier V, Baudin M, Orain G, Lemarquand A, Parisi L (2010b). Integrating scab control methods with partial effects in apple orchards: the association of cultivar resistance, sanitation and reduced fungicide schedules. *IOBC wprs Bulletin*, 54, 525-528.
- [8] Gamerman D., Lopes H.F. (2006) *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC
- [9] Gadoury D.M., MacHardy W.E. (1986). Forecasting ascospore dose of *Venturia inaequalis* in commercial apple orchards. *Phytopathology*, 76, 1128.
- [10] Gelman A., Rubin D.B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, 7, 457-511.
- [11] Geyer C.J. (1992) Practical Markov Chain Monte Carlo. *Statistical Science*, 7, 4, 473-483.
- [12] Gibson G.J. (1997) Markov Chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Applied Statistics*, 46, 215-233.
- [13] Guyon X., Pumo B. (2007) Space-time estimation of a particle system model. *Statistics*, 41 (5), 395-407.
- [14] Illian J., Penttinen A., Stoyan H., Stoyan D. (2008) *Statistical analysis and modelling of spatial point patterns*. Wiley, Chichester.
- [15] Johnson N.L., Kotz S., Balakrishnan N. (1995) *Continuous Univariate Distributions, Volume 2*. Wiley-Interscience
- [16] Lateur, M., Wagemans, C. and Populer, C. (1998). Evaluation of tree genetic resources as sources of polygenic scab resistance in an apple breeding programme. *Acta Horticulturae*, 484, 35-42.
- [17] MacHardy (1996) *Apple Scab : Biology, Epidemiology and Management*. The American Phytopathological Society (APS Press)
- [18] Madden L.V., Hughes G., Van Den Bosch F. (2007) *The study of Plant Disease Epidemics*. The American Phytopathological Society (APS Press).
- [19] Mills W.D., Laplante A.A. (1951) Diseases and insects in the orchard. *Cornell Extension Bulletin*, 711.
- [20] Mollison D. (1977) Spatial contact models for ecological and epidemic spread, with discussion, *J. of Royal Statistical society B*, 39, 283-326.
- [21] Olivier J.M. (1986) La tavelure du pommier, conduite d'une protection raisonnée. *Adalia* 1, 3-19.
- [22] Rasmussen J.G., Moller J., Aukema B.H., Raffa K.F., Zhu J. (2007) Continuous time modelling of dynamical spatial lattice data observed at sparsely distributed times. *J. of Royal Statistical society B*, 69, 4, 701-713.
- [23] Robert C.P., Casella G. (2004) *Monte Carlo Statistical Methods*. Springer
- [24] Roberts G.O., Gelman A., Gilks W.R. (1997) Weak Convergence and Optimal Scaling of Random Walk Metropolis Algorithms. *The annals of applied probability*, 7, 110-120.

- [25] Soubeyrand S., Laine A.-L., Hanski I., Penttinen A. (2009) Spatiotemporal Structure of Host-Pathogen Interactions in a Metapopulation. *The American Naturalist*, 174, 308-320.
- [26] Verhulst P.-F. Recherches mathématiques sur la loi d'accroissement de la population, *Nouv. mém. de l'Academie Royale des Sci. et Belles-Lettres de Bruxelles*, 18, 1-41, 1845.
- [27] Zhu J., Rasmussen J. G., Moller J., Aukema B., Raffa K. F. (2008) Spatial-temporal modeling of forest gaps generated by colonization from below- and above-ground bark beetle species. *J. of the American Statistical Association*, 103, 481, 162-177.