Modelling Spread of Diseases Using a Survival Analysis Technique

Evans Gouno
Université de Bretagne Sud, Campus de Tohannic, BP 573-56017, Vannes, France

Abstract
We propose a model to describe the spread of a disease among individuals regarded as fixed. The approach relies on a survival analysis technique working out times to infection. We reformulate the force of infection and introduce an infection factor referring to proportional hazard models. Properties of the MLE of the model parameters are studied. Results on real data are displayed and a simulation study is conducted.

Keywords: Epidemiology; Spatio-temporal model; Parasite diseases; Rate of infection

Introduction
The spatial and temporal progress of a disease in a field of non-moving individuals is an important issue which should be understood, in particular to set out control strategies. Models to describe the transmission are usually formulated using deterministic equations [1,2] but spatiotemporal stochastic models are also available [3-5]. The work presented here is motivated by an agricultural issue concerning sugarcane which can be infected with a yellowing and stunting disease called the sugarcane yellow leaf syndrome. The causal agent sugarcane yellow leaf virus (ScYLV) is transmitted by the aphid melanaphis sacchari. It is well-known that virus-free plants are quickly infected due to proximity to other infected plants. We consider that the infection rate of susceptible units at a given time depends on the distance to the infected areas. This question has been investigated by many authors [6-8]. Refer to Shaw [9] for a review of the application of spatiotemporal stochastic models in plant pathology.

Here, we have developed an approach based on survival analysis techniques by considering times to infection and introducing an infection factor to characterize the mechanisms which underlie the spread of the disease.

In section “Data”, we describe the nature of the data. Section “The model” presents the model. In section “Maximum likelihood estimation” we give a method to estimate the model parameters. Section “Applications” is devoted to application to real world data and simulation. Some concluding remarks are given in section “Concluding remarks”.

Data
We consider data of the following form: n units occupying the vertices of a finite, two-dimensional rectangular lattice L are observed. Each unit can be labeled by the vertex co-ordinates x. At set dates \( \tau_j \), \( j = 1, \ldots, m \), infected units are recorded (\( \tau_x = 0 \)).

For a unit \( x \in L \), the observation is a \( m \)-dimensional vector \( \delta_x = (\delta_{x,1}, \ldots, \delta_{x,m}) \), where for \( j = 1, \ldots, m \),

\[
\delta_{x,j} = \begin{cases} 
1 & \text{if } x \text{ is in the infected state between } [\tau_{j-1}, \tau_j] \\
0 & \text{if } x \text{ is the non-infected state between } [\tau_{j-1}, \tau_j].
\end{cases}
\]

An example of such data is given by the recording of the spread of ScYLV in a sugarcane field with 97 rows and 17 columns. The infected plants are recorded after 6, 10, 14, 19 and 23 weeks. The distance between rows is 0.5 m, and between columns 1.5 m. The numbers of infected plants are successively 6, 24, 68, 205 and 292. Figure 1 gives maps of the spread. Another example is given by Marcus et al. [7]. The spread of the citrus tristeza virus (CTV) in a citrus orchard is observed. The plants are arranged in a two-dimensional rectangular lattice with an inter-row distance of 5-6 m and a between column distance of 4 m. We have a total of 1008 units. 131 trees were recorded as infected in 1981 and 45 newly infected trees appear during the subsequent year. The map of the 176 infected trees is presented in figure 2.

In a series of papers, Gibson [3,4,10] and coworkers analyze these data and obtain accurate parameters values for the spread of CTV infection with a highly complex procedure. We suggest a simpler approach and consider more than two sequences of data in order to handle ScYLV data.

*Corresponding author: Université de Bretagne Sud, Campus de Tohannic, BP 573-56017, Vannes, France, Tel: +33(0)297017214; Fax: +33(0)297017071; E-mail: evans.gouno@univ-ubs.fr

Received January 10, 2011; Accepted March 18, 2011; Published March 22, 2011


Copyright: © 2011 Gouno E. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
The Model

Let us denote by \( T \), the time to infection for unit \( x \). We define the infection rate by analogy to the hazard function as

\[
\lim_{dt \to 0} \Pr(T_x \leq t + dt \mid T_x > t) = \frac{\lambda_x(t)}{dt}
\]

(1)

Thus \( \lambda_x(t) \) is the probability for a unit in position \( x \) to be infected in a small interval \( (t, t + dt) \) given that the unit was non-infected before \( t \). It is the probability of instantaneous infection or the infection hazard function.

We assume that:

\[
\lambda_x(t) = \phi_x(t)\lambda_0,
\]

(2)

where \( \lambda_0 \) is the baseline rate of transmission and where \( \phi_x \) is the infection factor or force of infection at time \( t \). This factor acts as the acceleration factor in reliability [11]. Here, we consider that the mechanism of transmission relies on contact, that is to say proximity to infected areas; function \( \phi_x \) depends on a distance \( d \) to infected area at time \( t \). Many choices of distance and many forms for \( \phi_x \) can be considered. A requirement is that \( \phi_x \) should be a decreasing function of the distance to infected areas. The closer a unit \( x \) is to infected units, the greater \( \phi_x(t) \) should be.

We have selected the following expression:

\[
\phi_x(t) = \exp \left\{-\gamma \inf_{y \in I_t} \frac{d(x,y)}{\phi_x} \right\}
\]

(3)

where \( I_t \) is the set of infected items at time \( t \) and \( d \) is the Euclidian distance. Combining (2) and (3) we obtain a model which can be viewed as the well known proportional hazards model introduced by Cox [12] where the covariate for a given items is characterized by the distance to infected areas. Thus in our model we have only one covariate which is time dependent and the baseline hazards function is assumed to be constant [13].

Note that one can add some more covariates (for e.g. species) depending on the purpose of the study.

Figure 3 displays the contamination factor for different values for \( \gamma \). One can see how the contamination factor behaves depending on parameter \( \gamma \) value for a same distance.

As mentioned in section “Data”, the data collected are usually grouped and the exact times to infection are not available. The inference relies on ‘snapshots’ of the epidemic at different times \( \tau_1, \ldots, \tau_m \). For each ‘snapshot’ the infection factor for unit \( x \) is computed as \( \phi_{x,j} = \exp \{- \gamma \inf_{y \in I_j} d(x,y) \} \) with \( I_j \) as the set of infected units at time \( \tau_j \).

For any unit \( x \), the infection rate at time \( t \) is

\[
\lambda_x(t) = \sum_{j=1}^{m} \phi_{x,j} \lambda_0 \mathbf{1}_{[\tau_j, \tau_{j+1})}(t).
\]

(4)

Thus \( \lambda_x(t) \) is a stepwise function where each step is a proportional hazards model [14].

Given expression (4), we express the probability for the time to infection to be greater than \( t \) as

\[
P(T_x > t) = \exp\left[-\sum_{j=1}^{m} \lambda_x(t) dt\right] \prod_{j=1}^{m} e^{\phi_{x,j} \lambda_0 \mathbf{1}_{[\tau_j, \tau_{j+1})}},
\]

and the probability density function of \( T_x \) is

\[
f(t) = \phi_{x,j} e^{\phi_{x,j} \lambda_0 \mathbf{1}_{[\tau_j, \tau_{j+1})}} \prod_{j=1}^{m} e^{-\phi_{x,j} \lambda_0 \mathbf{1}_{[\tau_j, \tau_{j+1})}}, t \in [\tau_{j-1}, \tau_j],
\]

In the following we assume that the times to contamination are independent and we propose a maximum likelihood method to estimate the parameters \((\lambda_0, \gamma)\) of the model using approximations of the times to infection.

Let us remark that because of the independence assumption it is not possible to consider correlation between pairs of observation in the classical way (that is to say computing correlation coefficient). But the link between times to infection of two non-infected units \( x \) and \( y \) can be considered relying on the proportional hazard formulation which allows to write: \( \lambda_x(t) = \phi_x(t)/\phi_y(t) \) and the ratio \( \phi_x(t)/\phi_y(t) \) measures in a sense the relationship between \( x \) and \( y \) at time \( t \).

Maximum Likelihood Estimation

Citrus tristeza virus

Before investigating the general case we consider the situation...
described by Marcus et al [7]. In this case, we only have two ‘snapshots’.
Let $\delta_x = 1$ if the tree is infected and $\delta_x = 0$ if it is not.

The likelihood is:
$$L(\lambda, \gamma) = \prod_{x \in \mathcal{X}} (\lambda_0 \phi_x \exp \{-\lambda_0 \phi_x \Delta / 2\} \exp \{-\lambda_0 \phi_x \Delta\})^{-\delta_x},$$
where $\phi_x = e^{-d_x}$ with $d_x$ the distance to the closest infected tree and $\Delta$ the difference between the data of inspections.

Let $k = \sum_{x \in \mathcal{X}} \delta_x$, the loglikelihood is expressed as:
$$\log L(\lambda, \gamma) = k \log \lambda_0 - \gamma \sum_{x \in \mathcal{X}} \delta_x d_x - \lambda_0 \sum_{x \in \mathcal{X}} d_x (1 - \delta_x/2) e^{-d_x},$$
and the likelihood equations are:
$$\begin{cases}
\sum_{x \in \mathcal{X}} (1 - \delta_x/2) e^{-d_x} = 0 \\
- \sum_{x \in \mathcal{X}} \delta_x d_x + \lambda_0 \sum_{x \in \mathcal{X}} d_x (1 - \delta_x/2) e^{-d_x} = 0
\end{cases}$$

A Newton-Raphson method is implemented to obtain the solution to these equations. We compute $\hat{\lambda}_0 = 0.142$ and $\hat{\gamma} = 4.186 \times 10^{-3}$.

Figure 4 shows epidemics simulated with the model. These images can be compared with those displayed by Gibson and Austin [3].

Simulating 10000 snapshots, the bias is $0.308 \times 10^{-4}$ for $\lambda_0$ and $0.666 \times 10^{-3}$ for $\gamma$. The mean squared error is $2.225 \times 10^{-6}$ for $\lambda_0$ and $1.707 \times 10^{-3}$ for $\gamma$.

The general case

Let us now consider the situation described in figure 1. Let $\Delta_i = \tau_{i+1} - \tau_i$. At time $\tau_{i+1}$, some units are already infected and they have no contribution to the likelihood. Some non-infected units at this time will be infected before $\tau_{i+1}$. Some others will remain non-infected at $\tau_{i+1}$. If unit $x$ is infected in $[\tau_{i+1}, \tau_i]$, we have $\delta_x = -\delta_{x-1} = 1$ and we assume that infection occurred at time $\Delta_i/2 = (\tau_{i+1} + \tau_i)/2$. In this case, the contribution to the likelihood is: $\lambda_0 \phi_x \exp \{-\lambda_0 \phi_x \Delta/2\}$. If unit $x$ is not in the state infected in this interval, $\delta_x = 0$ and the contribution to the likelihood is $\exp \{-\lambda_0 \phi_x \Delta/2\}$. $\delta_x = -\delta_{x-1}$ represents the newly infected units between $[\tau_{i+1}, \tau_i]$.

The likelihood is then:
$$L(\lambda, \gamma) = \prod_{i=1}^{k} \prod_{x \in \mathcal{X}} (\lambda_0 \phi_x \exp \{-\lambda_0 \phi_x \Delta_i/2\} \exp \{-\lambda_0 \phi_x \Delta_i\})^{-\delta_x}. (5)$$

where $\Delta_i = 0$

Computing the log-likelihood leads to:
$$\log L(\lambda, \gamma) = k \log \lambda_0 + \sum_{i=1}^{k} (\delta_{x_{i+1}} - \delta_{x_i}) \log(\phi_x)$$
$$- \lambda_0 \sum_{i=1}^{k} \phi_x \{1 - (\delta_{x_{i+1}} + \delta_{x_i})/2\} \Delta_i$$
where $k = \sum_{x \in \mathcal{X}} (\delta_{x_{i+1}} - \delta_{x_i})$

This log-likelihood is concave (see proof in annex). Thus there is a unique maximum likelihood estimate ($\hat{\lambda}_0, \hat{\gamma}$). Since the likelihood equations have a unique solution, then ($\hat{\lambda}_0, \hat{\gamma}$) is consistent, asymptotically normal and efficient [15].

We use a Newton-Raphson algorithm to obtain the estimates. A simulation study and an application on ScVL data are conducted in the following section.
Applications

Applying the method to the ScVLC data described in the introduction, we obtain: \( \hat{\gamma} = 0.23384 \) and \( \hat{\lambda_0} = 0.02224 \).

We applied the method on simulated data. We consider a lattice with 50 rows and 50 columns. The inter-row distance is equal to the inter-column distance: 1 m. We set 4 infected units at the initial time and observed the spread in successive windows of 4 units of time length for parameters values \( \hat{\lambda_0} = 0.159 \) and \( \gamma = 0.34 \). We generated data using the following scheme:

For \( t = 4, 8, 12 \) and for each non-infected units \( x \) at \( t = 4 \)
1. use equation (3) to compute \( \phi(t) \), the infection factor,
2. draw \( z \) from an exponential distribution with parameter \( \phi(t) \lambda_0 \),
3. if \( z < t \), set \( x \) infected.

Four different locations of the initial infected units are investigated. Note that dimension of the lattice, number and position of the infected units at the initial time was chosen arbitrarily as the values for \( \lambda_0 \) and \( \gamma \). Each row in figure 5 shows sequences of images for a single data set.

Table 1 gives the mean of the maximum likelihood estimates, the bias and the root mean square error (RMSE), for a large number of repeated data sets generated with the previous setting. The results surmise that the estimators are asymptotically unbiased and consistent in mean square.

Note that confidence intervals for \( \gamma \) and \( \lambda_0 \) can be obtained using properties of MLE.

Concluding Remarks

We have suggested a simple approach to model the spread of a disease in a field relying on survival analysis methods dealing with approximation of times to infection. This approach allows for further developments. For example, tests of hypothesis can be investigated to compare the spread of different diseases or spreading in different places depending on a distance to infected areas, but other factors can be incorporated in the model and a Bayesian development could be a possible direction as practitioners might have prior information on the propagation mechanism.

Annex: Existence and uniqueness of the likelihood estimates

Let \( \xi_{x,i}(t) = \phi(t) [1 - (\delta_{x,i} + \delta_{x,i-1})] \Delta_t \). We denote: \( \sum_{x_{i-1}}^{x} = \sum_{x_{i+1}}^{x} \), to lighten the notations. \( d_{x,i} \) is the distance for unit \( x \) to the closest infected unit at time \( \tau_c \).

The first derivatives of the log-likelihood are:

\[
\frac{\partial}{\partial \lambda_0} \log L(\hat{\lambda_0}, \gamma) = \frac{k}{\hat{\lambda_0}} - \sum_{x,i} \xi_{x,i}(\gamma)
\]

The likelihood equation system is then equivalent to:

\[
\begin{align*}
\lambda_0 &= k / \sum_{x,i} \xi_{x,i}(\gamma) \\
\phi(\gamma) &= 0
\end{align*}
\]

with \( \varphi(\gamma) = -\sum_{x,i} (\delta_{x,i} - \delta_{x,i-1}) d_{x,i} + k \sum_{x,i} \xi_{x,i}(\gamma) / \sum_{x,i} \xi_{x,i}(\gamma) \).

\( \varphi(\gamma) \) is a decreasing function. Indeed, the derivative of \( \varphi(\gamma) \) is:

\[
\varphi'(\gamma) = k \left[ \sum_{x,i} d_{x,i} \xi_{x,i}(\gamma) \sum_{x,i} \xi_{x,i}(\gamma) \left( \sum_{x,i} d_{x,i} \xi_{x,i}(\gamma) \right)^{-1} \right] \left( \sum_{x,i} \xi_{x,i}(\gamma) \right)^{-1}
\]

Since \( \xi_{x,i} > 0 \), we write:

\[
\left( \sum_{x,i} d_{x,i} \xi_{x,i}(\gamma) \right)^{-1} \left( \sum_{x,i} d_{x,i} \xi_{x,i}(\gamma) \right)^{-1}
\]

Applying the Cauchy-Schwarz inequality, we have \( \varphi'(\gamma) \leq 0 \), for all \( \gamma \).

Furthermore \( \lim_{\gamma \to 0} \xi_{x,i}(\gamma) = [1 - (\delta_{x,i} + \delta_{x,i-1})] / 2 \Delta_t = c_{x,i} \).

Then

\[
\lim_{\gamma \to 0} \varphi(\gamma) = \left( \sum_{x,i} (\delta_{x,i} - \delta_{x,i-1}) \right) \left( \sum_{x,i} d_{x,i} (c_{x,i} / \sum_{x,i} c_{x,i}) - d_{x,i} \right)
\]

(8) is greater than \( \sum_{x,i} (\delta_{x,i} - \delta_{x,i-1}) (d_{x,i} - d_{x,i-1}) \) where \( d_{x,i} \) is the maximum distance between two units. Thus \( \lim_{\gamma \to 0} \varphi(\gamma) \) is positive.

Since \( \xi_{x,i} < c_{x,i} \), we have: \( \sum_{x,i} \xi_{x,i} / \sum_{x,i} c_{x,i} < \sum_{x,i} \xi_{x,i} / \sum_{x,i} c_{x,i} \) and

\[
\varphi(\gamma) < \sum_{x,i} (\delta_{x,i} - \delta_{x,i-1}) \left( \sum_{x,i} d_{x,i} (c_{x,i} / \sum_{x,i} c_{x,i}) - d_{x,i} \right)
\]

which prove \( \lim_{\gamma \to 0} \varphi(\gamma) < 0 \) since \( \lim_{\gamma \to 0} \xi_{x,i} = 0 \).

It follows that \( \varphi(\gamma) = 0 \) has unique solution.

Acknowledgements

The author would like to thank the referees for their useful comments.

References


