Journal of Biometrics & Biostatistics

Gouno J Biomet Biostat 2011, 2:2
DOI: 10.4172/2155-6180.1000113

Research Article Open Access

Modelling Spread of Diseases Using a Survival Analysis Technique

Evans Gouno

Universitè 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 τ_j, j= 1,…,m, infected units are recorded ($\tau_0 = 0$).

For a unit $x \in$ L, the observation is a m-dimensional vector $\delta_x = (\delta_{x,1}\ \delta_{x,2}, \ldots, \delta_{x,m})$, 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 finite 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.

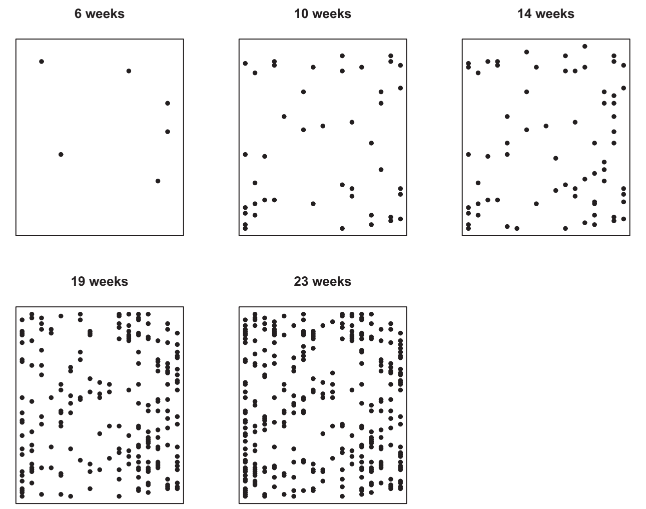

Figure 1: Data from CIRAD -Guadeloupe (FWI) Maps of newly infected plants at successive times in weeks.

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:** Universitè 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

Citation: Gouno E (2011) Modelling Spread of Diseases Using a Survival Analysis Technique. J Biomet Biostat 2:113. doi:10.4172/2155-6180.1000113

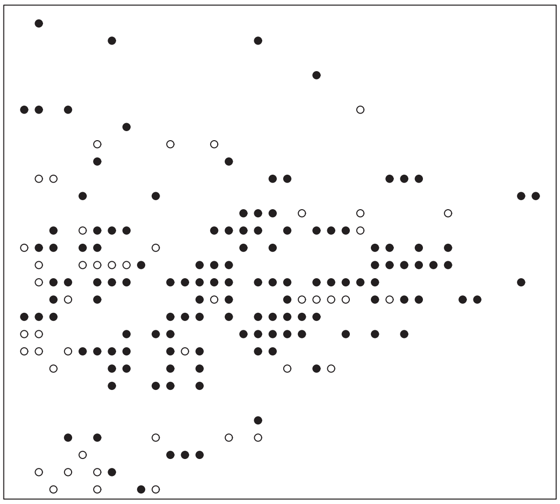

Figure 2: Location of trees infected by citrus tristeza virus. Black points indicate infected trees in 1981, and white points indicate trees discovered as infected in 1982.

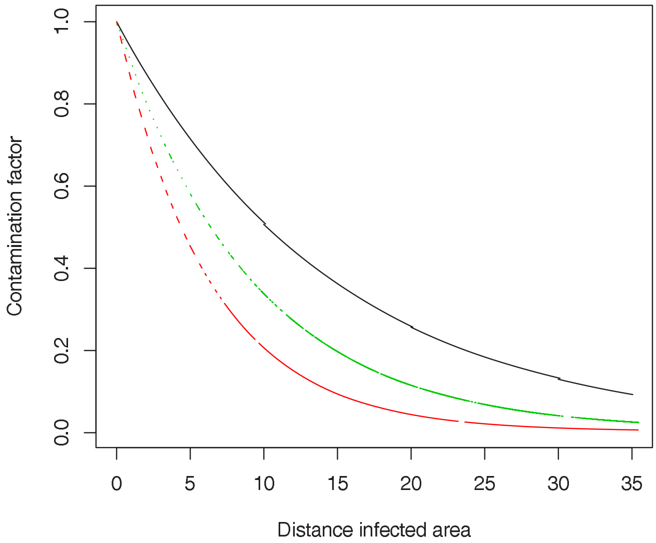

Figure 3: Representation of infection factor versus distance to infected area for = 0.068 (solid), 0.11 (dots) and 1.3 (dashes).

The Model

Let us denote by T_x the time to infection for unit x. We define the *infection rate* by analogy to the hazard function as

$$\lim_{dt\to 0}\frac{\Pr(T_x \le t+dt \mid T_x > t)}{dt} = \lambda_x(t) \tag{1}$$

Thus $\lambda_x(t)$ dt 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:

$$\lambda_x(t) = \phi_x(t)\lambda_0, \tag{2}$$

where λ_0 is the baseline rate of transmission and where ϕ_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 ϕ_x depends on a distance d to infected area at time t. Many choices of distance and many forms for ϕ_x can be considered. A requirement is that ϕ_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:

$$\phi_x(t) = \exp\left\{-\gamma \inf_{y\in I_t} d(x,y)\right\} \tag{3}$$

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 infection factor for different values for γ. One can see how the contamination factor behaves depending on parameter γ value for a same distance.

As mentionned 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 $\phi_{x,j} = \exp\{-\gamma \inf_{y\varepsilon Ij} d(x,y)\}$ with I_j as the set of infected units at time τ_j.

For any unit x, the infection rate at time t is:

$$\lambda_x(t) = \sum_{j=1}^{m} \phi_{xj}\lambda_0 1_{[\tau_{j-1,}\tau_j]}(t). \tag{4}$$

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

$$P(T_x > t) = \exp\{-\int_0^t \lambda_x(t)dt\} = e^{-\phi_{x,j}\lambda_0(t-\tau_{j-1})}\prod_{i=1}^{j-1} e^{-\phi_{x,i}\lambda_0\Delta_i}, t\in[\tau_{j-1},\tau_j],$$

and the probability density function of T_x is

$$f(t) = \phi_{x,j}\lambda_0 e^{-\phi_{x,j}\lambda_0(t-\tau_{j-1})}\prod_{i=1}^{j-1} e^{-\phi_{x,i}\lambda_0\Delta_i}, t\in[\tau_{j-1},\tau_j],$$

In the following we assume that the times to contamination are independent and we propose a maximum likelihood method to estimate the parameters (λ_0,γ) 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: $\lambda_x(t) = \phi_x(t)/\ \phi_y(t)\ \lambda_y(t)$ 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:

$$L(\lambda_0,\gamma) = \prod_{x\in\chi}[\lambda_0\phi_x \exp\{-\lambda_0\phi_x\Delta/2\}]^{\delta_x}[\exp\{-\lambda_0\phi_x\Delta\}]^{1-\delta_x},$$

where $\phi_x = \mathrm{e}^{-\lambda d_x}$ with d_x the distance to the closest infected tree and Δ the difference between the data of inspections.

Let $k = \sum_{x\in\chi}\delta_x$, the loglikelihood is expressed as:

$$\log L(\lambda_0,\gamma) = k\log\lambda_0 - \gamma\sum_{x\in\chi}\delta_x d_x - \lambda_0\Delta\sum_{x\in\chi} d_x(1-\delta_x/2)\, e^{-\gamma d_x},$$

and the likelihood equations are:

$$\begin{cases} \dfrac{k}{\lambda_0} - \Delta\sum_{x\in\chi}(1-\delta_x/2)e^{-\gamma d_x} = 0 \\ -\sum_{x\in\chi}\delta_x d_x + \lambda_0\Delta\sum_{x\in\chi} d_x(1-\delta_x/2)\, e^{-\gamma d_x} = 0 \end{cases}$$

A Newton-Raphson method is implemented to obtain the solution to these equations. We compute: $\hat{\gamma} = 0.142$ and $\lambda_0 = 4.186\ 10^{-3}$.

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 $0.308\ 10^{-4}$ for λ_0 and $0.666\ 10^{-3}$ for γ. The mean squared error is 2.225 10-6 for λ_0 and $1.707\ 10^{-3}$ for γ.

The general case

Let us now consider the situation described in figure 1. Let $\Delta_i = \tau_i - \tau_{i-1}$. At time τ_{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 τ_i. Some others will remain non-infected at τ_i. If unit x is infected in $[\tau_{i-1},\tau_i]$, we have $\delta_{x,i}-\delta_{x,i-1} = 1$ 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: $\lambda_0\phi_{x,i} \exp\{-\lambda_0\phi_{x,i}\Delta_i/2\}$. If unit x is not in the state infected in this interval, $\delta_{x,i}=0$ and the contribution to the likelihood is $\exp\{-\lambda_0\phi_{x,i}\Delta_i\}$. $\delta_{x,i}-\delta_{x,i-1}$ represents the newly infected units between$[\tau_{i-1},\tau_i]$.

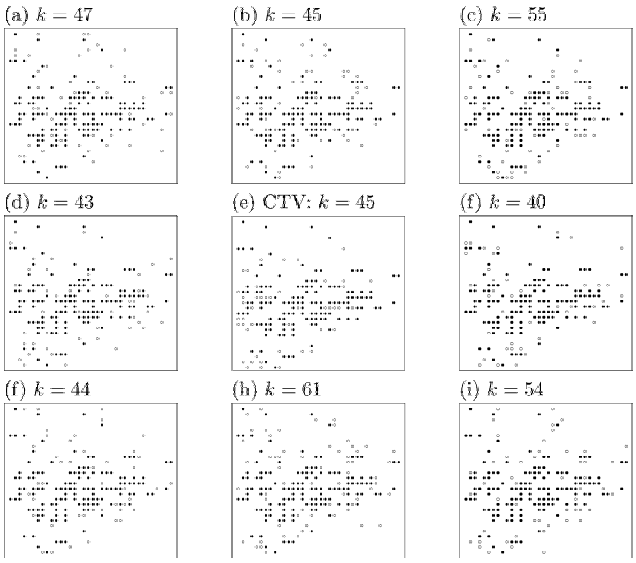

Figure 4: Snapshots of simulated infection using maximum likelihood estimates of λ_0 and γ. (e) is the observed CTV infection.

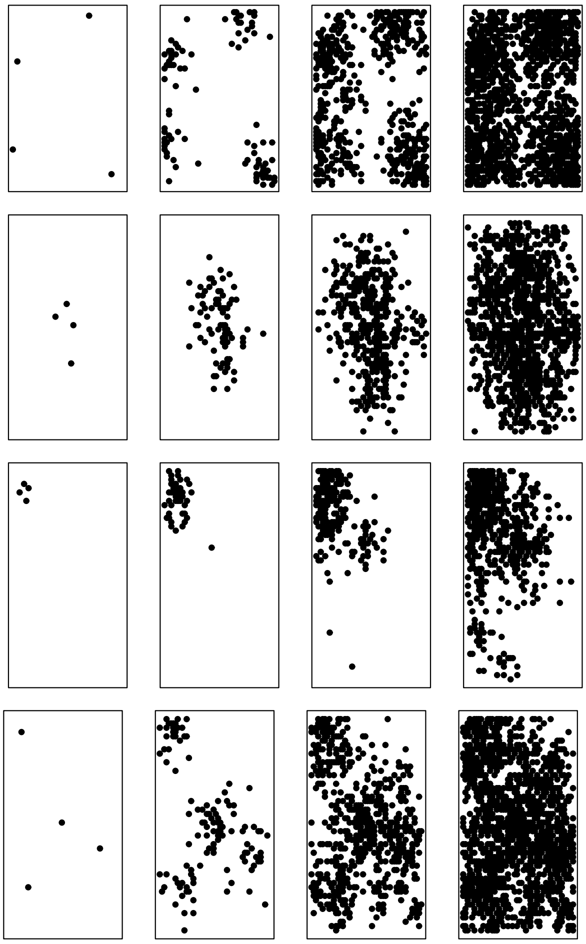

Figure 5: Simulations of the propagation of a disease with different initial positions of infected units with parameters values λ_0= 0.159 and γ = 0.34.

The likelihood is then:

$$L(\lambda_0,\gamma) = \prod_{x\in L}\prod_{i=1}^{m}[\lambda_0\phi_{x,i}\exp\{-\lambda_0\phi_{x,i}\Delta_i/2\}]^{(\delta_{x,i}-\delta_{x,i-1})}[\exp\{-\lambda_0\phi_{x,i}\Delta_i\}]^{1-\delta_{x,i}}. \quad (5)$$

where $\delta_{x,0} = 0$

Computing the log-likelihood leads to:

$$\log L(\lambda_0,\gamma) = k\log\lambda_0 + \sum_{x\in L}\sum_{i=1}^{m}(\delta_{x,i}-\delta_{x,i-1})\log(\phi_{x,i})$$

$$-\ \lambda_0\sum_{i=1}^{m}\phi_{x,i}[1-(\delta_{x,i}+\delta_{x,i-1})/2]\,\Delta_i$$

where $k = \sum_{x\in L}(\delta_{x,m}-\delta_{x,1})$

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 $\lambda_0 = 0.159$ 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 $\phi_x(t)$, the infection factor,
2. draw z from an exponential distribution with parameter $\phi_x(t)\lambda_0$,
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 λ_0 and γ. 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 γ and λ_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 to answer the question: are the mechanisms of transmission different? We have given some results on properties of estimators involved in the model. In this first approach, the infection factor is considered 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 $\xi_{x,i}(\gamma) = \phi_{x,i}\ [1 - (\delta_{x,i} + \delta_{x,i-1})/2]\ \Delta_i$. We denote: $\sum_{x\in\chi}\sum_{i=1}^{m} \equiv \sum_{x,i}$, to lighten the notations. $d_{x,i}$ is the distance for unit x to the closest infected unit at time τ_i.

The first derivatives of the log-likelihood are:

$$\frac{\partial}{\partial\lambda_0}\log L(\lambda_0,\gamma) = \frac{k}{\lambda_0} - \sum_{x,i}\xi_{x,i}(\gamma) \qquad (6)$$

	λ_0			γ		
	MLE	Biais	RMSE	*MLE*	Biais	RMSE
Border	0.156401	0.002599	0.007670	0.337438	0.002562	0.017201
Middle	0.147022	0,011978	0.031654	0.442686	0,102686	0.068031
Corner	0.156879	0,002121	0.011192	0.339409	0.000591	0.018539
Mixture	0.156005	0.002995	0.006838	0.337506	0.002494	0.013841

Table 1: MLE for simulated data with λ_0= 0:159 and γ = 0:34

$$\frac{\partial}{\partial\gamma}\log L(\lambda_0,\gamma) = -\sum_{x,i}(\delta_{x,i} - \delta_{x,i})\ d_{x,i-1} + \lambda_0\sum_{x,i}d_{x,i}\xi_{x,i}(\gamma) \qquad (7)$$

The likelihood equation system is then equivalent to:

$$\begin{cases} \lambda_0 = k/\sum_{x,i}\xi_{x,i}(\gamma) \\ \varphi(\gamma) = 0 \end{cases}$$

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

φ is a decreasing function. Indeed, the derivative of φ is:

$$\varphi'(\gamma) = k\left[-\sum_{x,i}d_{x,i}^2\ \xi_{x,i}(\gamma)\sum_{x,i}\xi_{x,i}(\gamma) + \left(\sum_{x,i}d_{x,i}\ \xi_{x,i}(\gamma)\right)^2\right]\Big/\left(\sum_{x,i}\xi_{x,i}(\gamma)\right)^2.$$

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

$$\left(\sum_{x,i}d_{x,i}\ \xi_{x,i}(\gamma)\right)^2\left(\sum_{x,i}d_{x,i}\sqrt{\xi_{x,i}(\gamma)}\sqrt{\xi_{x,i}(\gamma)}\right)^2$$

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

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

Then

$$\lim_{\gamma\to 0}\varphi(\gamma) = \sum_{x,i}(\delta_{x,i} - \delta_{x,i-1})\left[\sum_{x,j}d_{x,j}(c_{x,j}/\sum_{x,l}c_{x,l}) - d_{x,i}\right] \qquad (8)$$

(8) is greater than $\sum_{x,i}(\delta_{x,i} - \delta_{x,i-1})(d_{\max} - d_{x,i})$ where d_{max} 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: $\sum_{x,j}\xi_{x,j}/\sum_{x,l}\xi_{x,l} < \sum_{x,j}c_{x,j}/\sum_{x,l}\xi_{x,l}$ and

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

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

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

Acknowledgements

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

References

1. Dayananda P, Billard L, Chakraborty S (1995) Estimation of rate parameter and its relationship with latent and infectious periods in plant disease epidemics. Biometrics 51: 284-292.
2. Hasting A (1996) Models of spatial spread: is the theory complet? Ecology 77: 1675-1679.
3. Gibson G, Austin E (1996) Fitting and testing spatio-temporal stochastic models with applications in plant epidemiology. Plant Path 45: 172-184.
4. Gibson G (1997b) Markov chain monte carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. Appl Statist 46: 215-233.
5. Lewis MA (2000) Spread rate for a non linear stochastic invasion. J Math Biol 41: 430-454.
6. Hughes G, Madden LV (1993) Using the beta-binomial distribution to describe aggregated patterns of disease incidence. Phytopathology 83: 759-763.
7. Marcus R, Sveltana F, Talpaz H, Salomon R, Bar-Joseph M (1984) On the spatial distribution of citrus tristeza virus disease. Phytoparasitica 12: 45-52.

8. Smyth G, Chakraboty S, Clark RG, Pettitt A (1992) A stochastic model for anthracnose development in Stylosanthes scabra. Phytopathology 82: 1267-1272.
9. Shaw MW (1994) Modelling stochastic processes in plant pathology. Annu Rev Phytopathol.32: 523-544.
10. Gibson GJ (1997a) Investigating mechanisms of spatiotemporal epidemic spread using stochastic models. Phytopathology 87: 139-146.
11. Nelson W (1990) Accelerated Testing. John Wiley, New York, USA.
12. Cox DR (1972) Regression models and life-tables. J Roy Statist Soc B 34: 187-220.
13. Lawless JF (1982) Statistical Models and Methods for Lifetime Data. John Wiley & Sons, New York.
14. LeBlanc M, Crowley J (1995) Step-function covariate effects in proportional hazards model. Canadian Journal of Statistics 23: 109-129.
15. Lehmann EL (1991) Theory of Point Estimation. John Wiley & Sons, New York.
16. Chimard F, Vaillant J, Daugrois JH (2010) Modélisation de répartitions d’occurrences spatio-temporelles et épidémiologie végétale.