




An Individual-Based Spatial Epidemiological Model for the Spread of Plant Diseases

Martina CENDOYA , Ana NAVARRO-QUILES, Antonio LÓPEZ-QUÍLEZ, Antonio VICENT, and David CONESA

In the study of plant disease epidemics, the state of each individual in the population and their spatial location should be considered when modeling disease spread. We present a model to describe the spread of plant diseases, where the infection of a susceptible individual depends on the transmission rate of infected individuals and the spatial correlation. This latter is introduced through the Matérn correlation function, accounting for spatial dependence based on distance. Almond leaf scorch disease, caused by the bacterium *Xylella fastidiosa*, was used as a case study to test the behavior of the model parameters and the variability due to the characteristics and location of initial disease introduction using a proposed simulation algorithm. The greatest variability in the results depended on the range parameter of the Matérn correlation, i.e., the distance at which two observations can be considered spatially uncorrelated, and the initial introduction. The spatial distribution of individuals also had a strong influence on disease spread, highlighting that areas without trees acted as barriers when their extent was greater than the range parameter. It should be stressed that this individual-based model can be applied to other plant diseases, adapting the parameter values to their particular epidemiological characteristics.

Key Words: Individual-based models; Matérn correlation; Simulation; Spatial epidemic model.

1. INTRODUCTION

The development of new statistical methods and computational advances have made disease modeling a widely used tool to assess risks and predict the spread of diseases, which is useful in the design of effective strategies for their control (Keeling and Rohani 2008; Chen et al. 2014). Still, there are many factors involved in the spread of diseases that should be taken into account. Particularly, for the study of plant diseases, it is important to consider

M. Cendoya (✉) · A. Vicent, Centre de Protecció Vegetal i Biotecnologia, Institut Valencià d'Investigacions Agràries, Moncada, València, Spain (E-mail: cendoya_marmar@gva.es).

A. Navarro-Quiles · A. López-Quílez · D. Conesa, Departament d'Estadística i Investigació Operativa, Universitat de València, Burjassot, València, Spain.

© 2024 The Author(s)

Journal of Agricultural, Biological, and Environmental Statistics

<https://doi.org/10.1007/s13253-024-00604-2>

several aspects, including climatic conditions, transmission routes and human intervention in the spread of the pathogen, as well as the spatiotemporal changes in the distribution of susceptible host plants and the performance of methods for their detection and control (Cunniffe et al. 2015).

Compartmental epidemiological models have been used to study the epidemiology of diseases in humans and animals (Keeling and Rohani 2008), but also in plants (Gilligan 2002; Madden et al. 2007). These models assume that the population can be divided into different compartments according to the state of the disease. The simplest model is composed of three compartments dividing the population into three states, i.e., susceptible, infected and recovered individuals. These models are often built around differential equations, which describe the behavior of the disease at the population level (Kermack and McKendrick 1927). As reviewed by Kleczkowski et al. (2019), more complex models have been developed including different compartments, sub-classes of each compartment, and even stochastic approximations with discrete time-steps considering spatial heterogeneity in different ways. Parameter estimation in these models is usually performed, in the simplest way, using classical techniques. One of these techniques, known as least squares, consists of obtaining those parameters that minimize the distance between the model and the data (Matis and Hartley 1971; Chowell et al. 2009; Capaldi et al. 2012). However, there are currently other methodologies that are becoming increasingly important such as MCMC methods within the Bayesian framework (Gibson 1997; Gibson et al. 2004) and metaheuristic algorithms such as the genetic algorithm (GA) and particle swarm optimization (PSO) (Akman and Schaefer 2015; Akman et al. 2018).

A drawback of compartmental models based on differential equations is that it is not possible to identify the individuals fluctuating between the compartments. The rate of transmission is then proportional to the number of susceptible individuals, the proportion of infectious individuals and the rate of contacts among individuals (Kermack and McKendrick 1927; Kleczkowski et al. 2019). An alternative approach consists of using individual-based models, which make it possible to monitor the state of each individual and the interactions among them. Nevertheless, the number of iterations increases substantially with large population sizes and long time horizons, leading to high computational costs that can limit the implementation of these models (Keeling and Rohani 2008).

Modeling spread requires also a good knowledge of the spatiotemporal behavior of the disease. When including spatial structures in the models, it is usually assumed that interactions among individuals, and thus potential disease transmission, decrease with distance. However, long-distance spread due to human interventions or other factors cannot be ignored, as it plays an important role in shaping disease progress. It is also important to consider the spatial heterogeneity of populations since host aggregation and distance from the source of transmission can strongly influence model outcomes (Keeling and Rohani 2008).

The spatial component of plant disease spread has been approached from different methods, such as partial differential equations, distance class methods, spatial autocorrelation, or metapopulations (Madden et al. 2007; Meentemeyer et al. 2011). The scale and the type of spatial relationship in the models are often determined by the availability of data and the computational cost required for their analysis. Using a fine spatial scale can have a high

computational cost, but spatial effects cannot be detected when aggregating the data at a larger scale. For instance, with lattice (aggregated) data modeling, the neighborhood structure is an approximation of the real dynamics, thus increasing the speed of the simulations (Keeling and Rohani 2008). However, working at a finer scale (i.e., at the individual level of a population with individuals represented as discrete points in time and space) allows the spread of the disease to be described more accurately.

The spatiotemporal dynamics of an individual-based model can be described through the probability of a susceptible individual becoming infected. This probability is quantified as a function of the surrounding infectious population and the effect of distance between individuals on disease transmission, which is usually incorporated by different forms of kernels (e.g., Keeling et al. 2001; Deardon et al. 2010; Meentemeyer et al. 2011; Hyatt-Twynam et al. 2017).

The spatial dependence can be integrated into the model by means of a correlation function. Although several functions have been proposed to handle the spatial effect, the Matérn correlation function (Matérn 1986) is widely used in several fields due to its flexibility, since it encompasses several functions depending on the value of its smoothing parameter (Stein 1999; Guttorp and Gneiting 2006). Despite its advantages, to date, the Matérn correlation function has not been integrated into spatial individual-based models for plant disease spread.

The main objective of this study was to develop a spread model for plant diseases, (i) based on individuals, thus identifying their state at any given time, (ii) considering spatial dependence, and (iii) that was computationally efficient. For this purpose, from the available data, an optimal value for the disease transmission rate was obtained using the classical structure of a compartmental model, to which the spatial dependence was also included through the Matérn correlation function. Then, a spatially explicit algorithm for the disease spread at individual level was designed.

Almond leaf scorch disease (ALSD), caused by the bacterium *Xylella fastidiosa*, was used as a case study. In particular, disease spread in almond trees in an area of Alicante (Spain) was simulated using this individual-based model. The effects of the parameters on disease spread were evaluated, as well as the different types and locations of initial disease introduction. A user-friendly tool to visualize the results was also developed. Spatial individual-level prediction of disease spread provides a better understanding of the epidemiology, thus allowing the application of more targeted control measures and the optimization of resources.

2. EPIDEMIOLOGICAL MODEL

This section describes the proposed framework for modeling plant disease spread (i.e., spatiotemporal progression), based on the disease status of each individual and the influence of their spatial dependence. The proposed methodology is structured into two distinct phases. Firstly, we employ a compartmental model, considering all individuals at each moment in time. This model is governed by a parameter-dependent system of differential equations, determining the transitions of individuals among various states, which also includes spatial

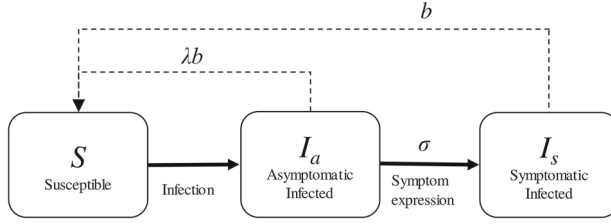


Figure 1. Compartments and dynamics of the epidemiological model. The solid arrows indicate movement of individuals between states. The population starts as susceptible (S), after infection goes through an asymptomatic infected (I_a) period and continues with the expression of symptoms, becoming symptomatic infected (I_s). The dashed arrows indicate the disease transmission (b) from infectious (I_a and I_s) to susceptible individuals, where b is given by the transmission rate and spatial correlation. λ is a transmission reduction parameter for I_a individuals.

dependence. Should real data be available, the parameters involved in both phases can be estimated using either classical deterministic techniques or stochastic methods, thereby acknowledging and accounting for the inherent uncertainty. Since this model only provides us with information at the population level, this concept is extended to a model based on individuals. We conclude the section by presenting an algorithm to simulate using our proposed model.

2.1. COMPARTMENTAL MODEL

Based on the structure of the compartmental model, individuals are classified according to their disease status as susceptible, i.e., non-infected individuals, or infected. Depending on the disease and the organism affected, the number of compartments and the transition between them can be highly variable. Models with more compartments may be closer to reality but also imply greater complexity in data fitting, require greater availability of information, and are more case-specific. Here, following a basic model with the usual states for most plant diseases, three possible disease states were defined for individuals: susceptible (S), those that are susceptible but are not infected; asymptomatic infected (I_a), those infected and infectious but without visible symptoms; and symptomatic infected (I_s), those infected and infectious with visible symptoms. Following the scheme in Fig. 1, infection of S individuals comes from both I_a and I_s . The disease transmission rate of infectious individuals, i.e., I_a and I_s , is defined by β . Since asymptomatic individuals have a lower pathogen concentration (European Food Safety Authority (EFSA) 2019), the transmission rate of I_a is reduced by the parameter λ . After an asymptomatic period, the transition from I_a to I_s , i.e., the time elapsed for the symptoms expression, is given by σ . The transmission rate and population size are assumed constant over time.

However, the spatial dependence of individuals and their spatial distribution can influence disease spread. This can be taken into account by incorporating any possible spatial dependency between susceptible and infectious individuals, such as a correlation function implying a distance-dependent decrease in disease transmission. In particular, we used the Matérn correlation function (Matérn 1986), a very flexible correlation family that generalizes many of the correlation functions widely used in spatial statistics. For two locations s_i and s_j , separated by a Euclidean distance $d_{ij} > 0$, the Matérn correlation is:

$$C(d_{ij}) = \frac{2^{1-\nu}}{\Gamma(\nu)} (\rho d_{ij})^\nu K_\nu(\rho d_{ij}), \quad (1)$$

where K_ν is the Bessel modified function of the second kind and order ν , which is the smoothness parameter of the function, and ρ is a scale parameter defined by the range parameter (r), considered as the distance at which two locations can be considered spatially uncorrelated. The Matérn class comprises different special cases with only two parameters, providing a wide range of forms of correlation functions, including the exponential correlation function when $\nu = 1/2$ (Handcock and Wallis 1994; Guttorp and Gneiting 2006). The literature contains different parameterizations of the Matérn correlation function which define the scale parameter in a variety of ways (see Handcock and Wallis 1994; Stein 1999; Diggle et al. 2003). We used the one described by Lindgren et al. (2011), which facilitates the parameters interpretation, where from the empirically derived definition $r = \sqrt{8\nu}/\rho$. Through this relation, r represents the distance at which the correlation is close to 0.1 for all ν .

The flow of individuals from one compartment to another is unidirectional and with no recovery option. Thus, as the number of infected individuals increases, the number of susceptible individuals decreases, with no possible return to the susceptible state. Under these assumptions, we can define a deterministic compartmental model that describes the disease progression by means of the solution of the following ordinary differential equations (ODEs):

$$\begin{aligned} \frac{dS}{dt} &= -(\lambda b I_a + b I_s) S, \\ \frac{dI_a}{dt} &= (\lambda b I_a + b I_s) S - \sigma I_a, \\ \frac{dI_s}{dt} &= \sigma I_a. \end{aligned} \quad (2)$$

The parameter b includes the transmission rate β and the spatial dependence as:

$$b = \beta \sum_{i \in \Omega_{S(t)}} \frac{1}{|\Omega_{S(t)}|} \sum_{j \in \{\Omega_{I_a(t)}, \Omega_{I_s(t)}\}} C_{ij}, \quad (3)$$

such that C_{ij} is the spatial correlation between individuals i and j , where $\Omega_{S(t)}$ is the set of susceptible individuals at time t and $\{\Omega_{I_a(t)}, \Omega_{I_s(t)}\}$ the set of infectious individuals, asymptomatic and symptomatic, at time t . Thus, b encompasses, together with the transmission rate β , the mean for the set of susceptible individuals i of the cumulative spatial correlation for each individual i to the infectious individuals j .

When real data are available, specifically including the number of susceptible and infected individuals, both symptomatic and asymptomatic, at different time points, it becomes feasible to estimate the parameters of the system, λ , β , σ , and the parameters of the Matérn correlation function. From a deterministic perspective, one approach to estimate these parameters is by minimizing the distances between the solution, which relies on the parameters, and the observed data at each time point (least square method). In this work, given the limited availability of data and that the main objective is to simulate the disease spread from the

origin, we focus on analyzing the impact of these parameters on the spread at individual level, where the system of equations presented is used as a basis for obtaining a reference value for the transmission rate.

2.2. SPATIAL INDIVIDUAL-BASED MODEL

The system of ODEs in Eq. (2) provides a simplified representation of disease progression at the population level. In contrast, individual-based models allow a more detailed analysis of disease spread, taking into account individual characteristics, interactions and spatial distribution. In individual-based models, the infection of a susceptible individual depends on the force of infection (φ_i), which incorporates the prevalence of infected individuals, i.e., the infected proportion of the total population, the transmission rate, and the spatial interaction between individuals (Keeling and Rohani 2008). In line with this and with the proposed compartmental model (Eq. 2), our proposal for the force of infection parameter φ_i , for a susceptible individual i at the time t , includes a dependence on the surrounding infected individuals through the Matérn correlation $C(d_{ij})$ for each pair of individuals i and j at a Euclidean distance d_{ij} :

$$\varphi_i(t) = \lambda\beta \sum_{j \in \{\Omega_{I_a(t)}\}} C(d_{ij}) + \beta \sum_{k \in \{\Omega_{I_s(t)}\}} C(d_{ik}), \quad (4)$$

where j and k represent each of the I_a and I_s individuals, respectively, of the infected population at time t (i.e., $\Omega_{I_a(t)}$ and $\Omega_{I_s(t)}$); $\beta > 0$ is the transmission rate of infection; and $0 \leq \lambda \leq 1$ represents a decrease in the transmission rate of I_a with respect to I_s .

Given this force of infection, the probability of a susceptible individual becoming infected at time t is $P_i(t) = 1 - \exp(-\varphi_i(t))$ (Keeling and Rohani 2008; Deardon et al. 2010).

2.3. SIMULATION ALGORITHM

An algorithm to simulate disease spread based on the previous model was implemented using the Python programming language (Python Software Foundation 2021). The Numba library (Lam et al. 2015) was also used due to its computational efficiency with high-dimensional data and its ability to run processes in parallel threads.

As shown in the diagram of the algorithm in Fig. 2, the landscape, i.e., the georeferenced locations of the individuals, must first be defined. The onset of the outbreak is initiated with an initial disease introduction I_a at $t = 0$. At each time t , S individuals are evaluated. For each S_i , φ_i is calculated based on the surrounding infected individuals at a distance less than $d_{\max} = 1.5r$, taking into account whether they are I_a or I_s following Eq. (4). This maximum distance to compute the spatial correlation allows the algorithm to be optimized, considering that at a distance greater than r the correlation between two locations is almost zero while capturing the tail of the correlation function. The infection of individual S_i is assessed by the discretization of the probability of infection P_i , through a random variable X_i from a Bernoulli distribution $\text{Be}(P_i)$, i.e., if $X_i = 1$ individual i becomes infected I_a , and if $X_i = 0$ individual i does not become infected.

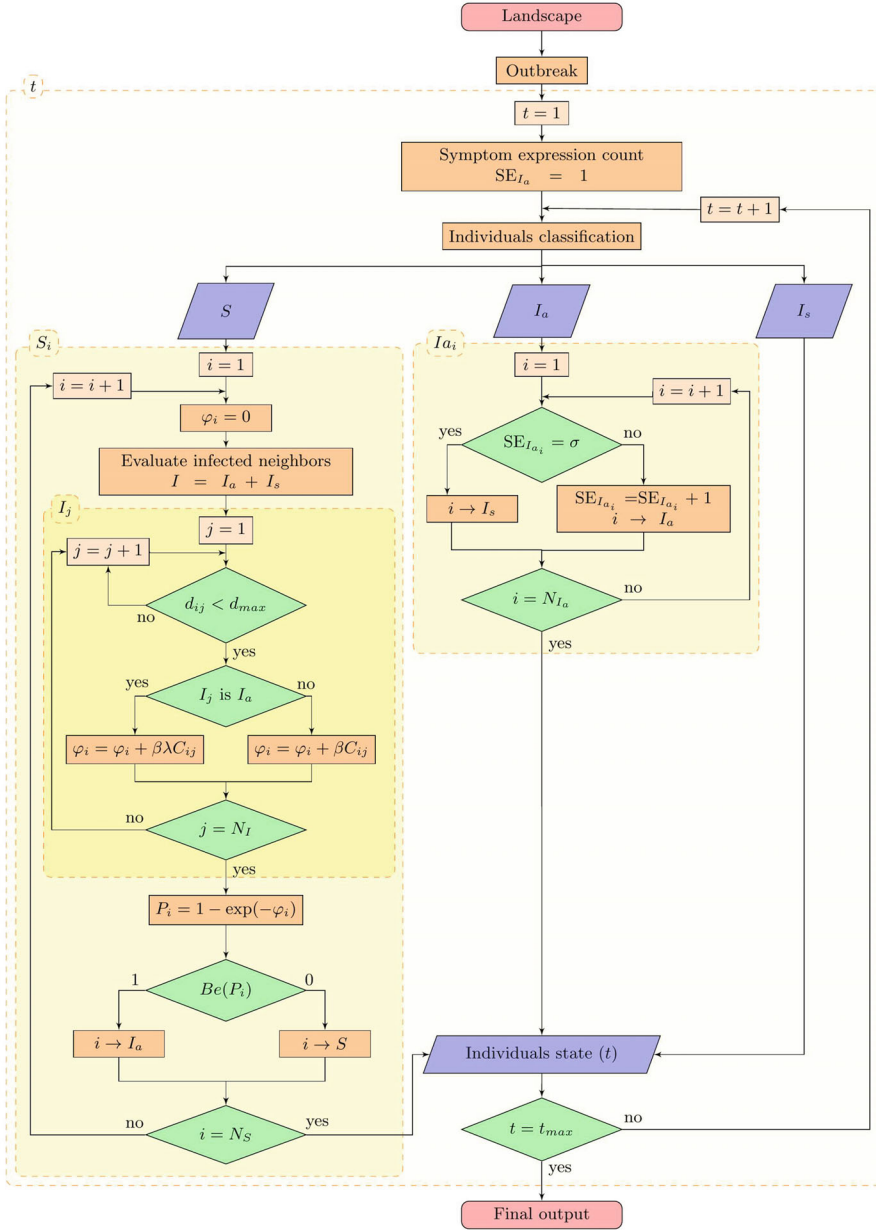


Figure 2. Diagram of the simulation algorithm. S and I indicate the susceptible and infected states, respectively, where I differentiates asymptomatic (I_a) and symptomatic (I_s). N_S , N_I and N_{I_a} are the total number of individuals in each corresponding state. d is the Euclidean distance and d_{max} is the maximum distance to evaluate neighbors. φ represents the force of infection, β is the transmission rate of infection, reduced by the λ parameter for I_a . C is the spatial correlation, P is the probability of infection, and $Be(P)$ denotes a random value of Bernoulli distribution with probability P . t represents the time and t_{max} the maximum time set for the simulation. The symptom expression SE counter starts when a susceptible individual becomes infected (I_a), so that when the given time for symptom expression (σ) is reached, I_a becomes I_s . The dashed line squares contain the loops for performed by the algorithm.

Likewise, for each t the I_a individuals are evaluated. For this purpose, a counter is initialized at $t = 1$ for the symptom expression of the individuals in this state. When the given value for the parameter σ is reached, individual i becomes I_s . Finally, the result of the state of all individuals for each t is obtained.

3. CASE STUDY

This section describes the rationale and the case study selected for the application of the individual-based spread model. We performed several simulations with the algorithm to assess disease spread. Simulations were performed in the study area to compare the disease spread in space and time under different scenarios and to evaluate the efficiency of the algorithm with a large number of individuals. To interactively visualize the results of all scenarios, a Shiny application (Chang et al. 2021) was developed using R software (R Core Team 2020). It is available at https://spatial-ibm.shinyapps.io/spread_results_app/.

3.1. RATIONALE

Almond leaf scorch disease (ALSD) is caused by the plant pathogenic bacterium *Xylella fastidiosa* (Wells et al. 1987). This pathogen inhabits the vascular tissues of the xylem, causing leaf scorch, wilt, dieback, and plant death. The bacterium is disseminated through insect vectors as well as propagating plant material (Almeida et al. 2005). Before its first detection in Europe in 2013, the pathogen was only known to be present in America (Saponari et al. 2013). After this first report in Italy, several outbreaks have been detected in Europe, potentially infecting more than 500 plant species (European Food Safety Authority (EFSA) 2020). In Alicante (Spain), *X. fastidiosa* was reported for the first time in 2017 affecting almond (*Prunus dulcis*) trees. A demarcated area was delimited based on the EU legislation (European Commission 2022), where intensive surveys and disease control measures are implemented by the Plant Health Authority in accordance with the Commission Implementing Regulation (EU) 2020/1201. As 90.3% of the positive samples for the presence of the bacterium were found in almond trees (Generalitat Valenciana (GVA) 2021), it is considered the main host of *X. fastidiosa* in this area and therefore is used as the susceptible population in our case study.

3.2. STUDY AREA, POPULATION AND PARAMETERS

The study area covered the *X. fastidiosa* infested area in Alicante. To simulate the spread of ALSD at the individual level, the Agricultural Plot Geographic Information System (SIGPAC) database and the grid used in the official surveillance program were merged to generate the georeferenced distribution of almond trees. From SIGPAC, we selected the plots identified as ‘nut trees’ and ‘nut trees with other associated crops’ (Ministerio de Agricultura, Pesca y Alimentación (MAPA) 2021). The study area was defined to the boundary of the grid cells with samples from the official surveillance program. As the locations of the trees were not available in the databases, the georeferenced distribution of almond trees was generated

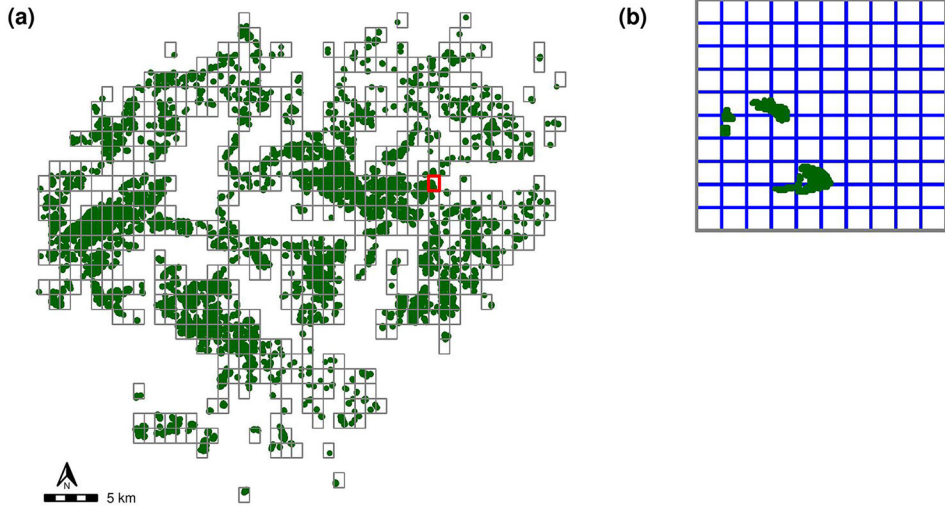


Figure 3. **a** Georeferenced distribution of almond trees (green) generated in the affected area of Alicante. The grid of cells aggregated to 1 km^2 is represented, where the cell marked in red has been zoomed in **b** with the corresponding grid of 1 ha cells used in the simulations, and the almond trees corresponding to these cells (green) (Color figure online).

with a $7 \times 7 \text{ m}$ tree row spacing, which is frequent in the traditional system of almond tree planting (Segura et al. 2018) in the selected SIGPAC plots. Our study area thus consisted of 282,041 almond trees (Fig. 3).

The effect of the model parameters (β , r and ν) on disease progression was compared by simulation. 100 I_a infected individuals were initially introduced and a 360-month time period was simulated. The disease spread was also evaluated according to the type of initial introduction. Specifically, three types of initial introduction were evaluated, referred to hereafter as random, at 5 foci and at one focus. In the case of random introduction, the initial 100 I_a individuals were randomly assigned, generating a more or less geographically dispersed distribution of infected individuals over the study area. For the initial introduction at one focus, an individual was randomly selected in the study area, defining the center of a circumference with a radius of 1 km where the remaining the individuals were then randomly selected. The same process was used for the introduction at 5 foci, initially selecting 5 individuals randomly from the population, and assigning the remaining in a radius of 1-km around them so that each focus consisted of 20 individuals I_a . In this way, each initial introduction type started with the same number of I_a with the different configurations relative to geographic location. In all cases, random selection was performed by non-replacement sampling where all individuals had the same probability of being selected.

Due to the relatively slow progression of the disease (Moralejo et al. 2020), simulations were performed at monthly time-steps to improve computational efficiency. Although molecular analyses indicate that the first outbreak of the disease in Alicante was decades ago (Landa et al. 2020; Moralejo et al. 2020), it was not detected until 2017. Therefore, data for estimating parameters describing the disease from its origin are scarce. Official surveillance data for *X. fastidiosa* from 2018 to 2021 were used to compare the prevalence observed with

Table 1. Average of the spatial correlations C obtained for the set of susceptible individuals with each of the values of the range parameter r of the Matérn correlation function, and the corresponding optimal transmission rate β values obtained by the sum of the squared error of the surveillance data and values predicted values by the compartmental model

r (m)	C	β (month ⁻¹)
250	2.12	0.014
500	3.15	0.01
750	4.19	0.007
1000	5.39	0.006

the results of the fitted compartmental model in Eq. (2) and obtain a reference value for β . The compartmental model was fitted for 30 years with an initial outbreak of 100 individuals I_a of the approximate almond tree population in the study area (282,041 almond trees). Since there was unavailable information on the asymptomatic period, the parameters related to a reduction in the transmission rate of I_a (λ) and the rate of transition from asymptomatic to symptomatic ($\sigma = 1/\text{asymptomatic period}$) were set according to information gathered from the literature as $\lambda = 0.015$ (White et al. 2020) and $\sigma = 1/(8 \text{ months})$ (European Food Safety Authority (EFSA) 2019). The spatial dependence introduced in the compartmental model was obtained from the spatial correlation through the Matérn function (Eq. 1, between each susceptible individual, i.e., negative sampling result, i.e., and each infected individual, i.e., with positive sampling result. The correlation average C of the set of individuals S was used from the accumulated correlations for each individual S following Eq. (3). The model was run with 100 values of β between 0.001 and 0.1 with C obtained given four values for the range parameter r of the Matérn correlation function: 250, 500, 750 and 1000 m (Table 1). The value of β that minimized the sum of the squared error of the infected individuals (I) predicted by the model and the observations was considered optimal. The sum of the squared error was given as $\sum_{i=1}^n (I(t_i, \beta) - I(t_i))^2$, where n is the total time t evaluated. For this purpose, the proportion of infected individuals predicted by the model for the last 4 years, i.e., $n = 4$, and the proportion of infected individuals accumulated from the samples in relation to the total number of samples for each year were used. As a result, the optimal value obtained for β was between 0.006 and 0.014 month⁻¹, with $r = 1000$ m and $r = 250$ m, respectively (Table 1). However, it must be taken into account that this optimization was only based on the data available for the last few years.

3.3. RESULTS OF DISEASE SPREAD SIMULATIONS

In the following, we present the results obtained when simulating with different parameter values and different initial introductions in order to evaluate: (i) the effect of the different parameters on the disease spread, (ii) the variability depending on the initial introduction, and (iii) the intrinsic variability of the model. Because of the large number of simulations and computational cost due to the large number of individuals, infection probabilities were calculated using a grid of 1 ha cells (Fig. 3). At each time-step, one individual S per cell was randomly selected and the probability of infection of this individual was assigned to

the other individuals S in the same cell. To test whether this approach produced biases in the results, we compared the results obtained with the simulation of disease spread by calculating the probability of infection of each individual S at each time t .

3.3.1. Effects of Parameters

Disease spread was simulated with the combination of different values for the parameters β , r and ν , with initial introduction of the infected individuals random, at 5 foci or at one focus. For each of these types of introduction, the location of the initial infected individuals was fixed at the same location. The disease spread simulations were performed with β values of 0.005, 0.015 and 0.03 month⁻¹, thus covering the range of optimal values of the transmission rate obtained from the surveillance data of *X. fastidiosa* in the study area with the compartmental model. Different values of the parameters of the Matérn r and ν correlation function also were tested. The values of r used, as for the β optimization, were 250, 500, 750 and 1000 m, representing a reasonable range of values for the spread distance by insect vectors. The smoothness parameter ν was set to 0.5, 1 and 1.5. The transmission rate reduction for I_a was set as $\lambda = 0.015$ and the time to symptom expression at 8 months for all simulations (European Food Safety Authority (EFSA) 2019).

The results were similar for the three values of ν , regardless of the other parameters (β and r) or the type of initial introduction (Fig. 4). The slight variations when the initial introduction was aggregated at 5 foci or at one focus can be attributed to the intrinsic variability of the simulations. Therefore, due to the lack of variation concerning this parameter, $\nu = 1$ was fixed in subsequent simulations.

With the same r and the same type of introduction, few differences were found when varying β . With the lowest β , the percentage of S decreased moderately slower. Despite this, after 360 months, the results were almost the same. The only notable difference after this time period was obtained with the introduction at one focus and $r = 750$ m, where the percentage of S was 36.24% higher with the lowest β compared to that obtained with the highest β , while in all other combinations this difference in the percentage of S did not exceed 7%.

Simulations with a large r value resulted in a higher infection rate. Nevertheless, the spread of the disease was different depending on the type of initial introduction. When the initial introduction was aggregated at one focus, with low values of r the decrease in the percentage of S individuals stabilizes in a relatively short period of time. However, with $r = 1000$ m it continues to decrease at different rates over time. This behavior was also observed when the initial introduction was aggregated at 5 foci. With the two lowest values of r , however, the number of susceptible individuals continues to decrease slowly without stabilizing, and with $r = 1000$ m the fluctuations in the infection rate were less pronounced. These variations were not observed with the random initial introduction, where the number of infected individuals almost stabilized at 96 months at about 2 and 0.6% of S with $r = 750$ m and $r = 1000$ m, respectively, while with the lowest values of r the disease continued to spread slowly.

Large differences were found depending on the type of initial disease introduction and on the value of r . While in the case of random introduction the percentage of S fell rapidly in

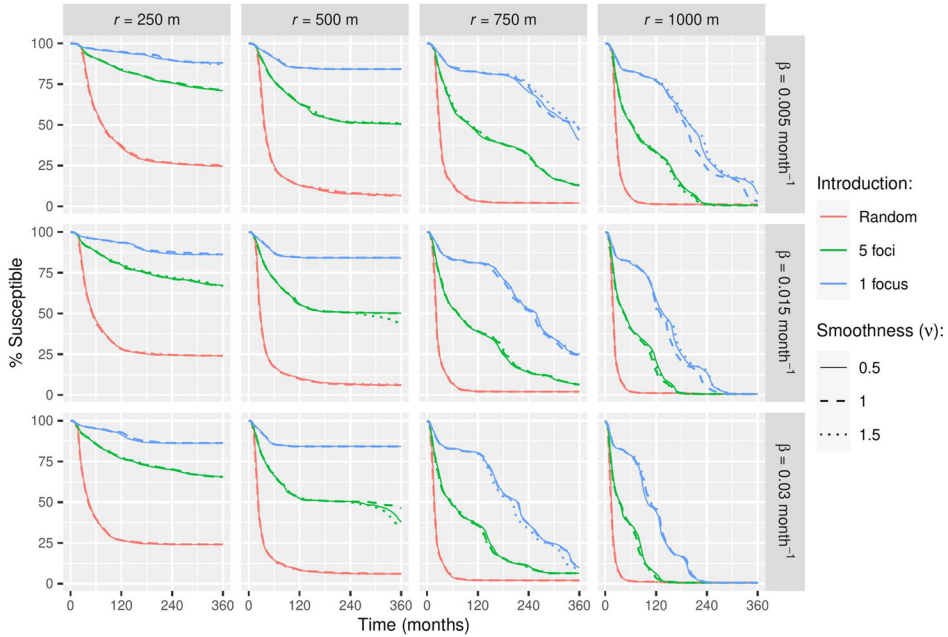


Figure 4. Percentage of susceptible individuals over time for the simulations with different values of the transmission rate (β), the parameters of the Matérn correlation function, range (r) and smoothness (ν) and the type of initial introduction (random, 5 foci and one focus).

the first few months, the decline was more gradual when introduction was aggregated, being slower the higher the aggregation, i.e., when the introduction was at one focus. Disease spread was limited in all cases by the value of r , so that at the highest value of r , regardless of β or introduction, after 360 months only 0.6% of S remained. At $r = 750$ m, disease spread was rapid with random introduction and stabilized after a few months at 2% of S , whereas with both aggregated introductions the decline continued without appearing to stabilize during this period. At the two lowest values of r (250 and 500 m), the percentage of S practically stabilized in all cases. The higher r and the less aggregated the introduction, the faster this stabilization point was reached and the lower the percentage of remaining S , i.e., the disease affected a greater number of individuals.

3.3.2. Variability Due to Initial Introduction

To study the variability due to the type and geographical location of the initial introduction, 10 simulations were performed for each type of introduction, random and aggregated at 5 foci and at one focus, with each combination of β and r parameter values. In each simulation, the location of the initially infected individuals was different, but in accordance with the type of introduction, choosing the introduction of the I_a as described above.

The dynamics of disease spread were similar to those observed previously in all cases (Fig. 5). However, while the variability between different introduction configurations was very low for random introduction, large differences were observed for aggregated introduction, depending on the location of the initial foci of infected individuals. The difference

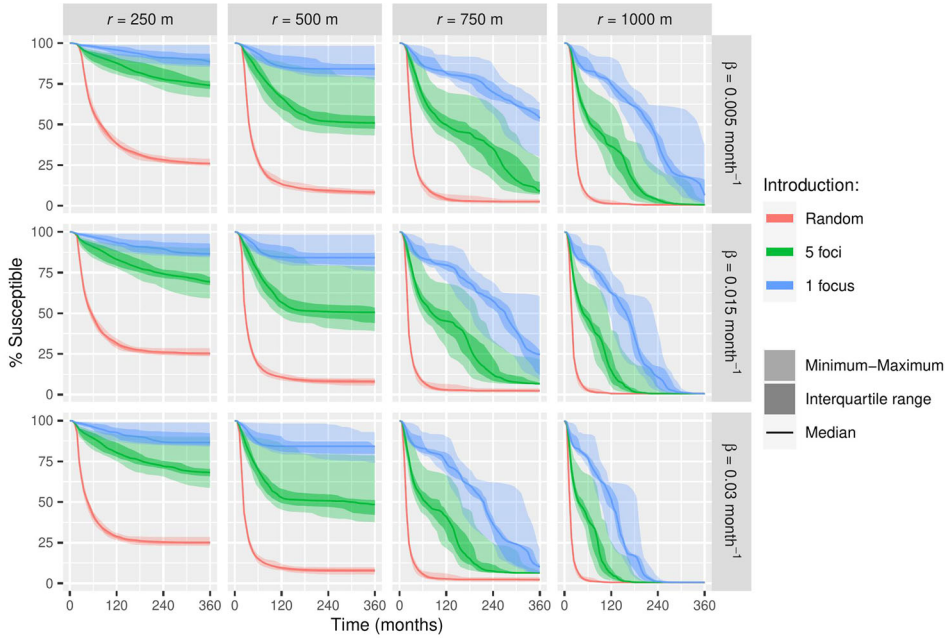


Figure 5. Minimum, maximum, interquartile range and median of the percentage of susceptible individuals over time of the 10 different geographical locations for each type of initial introduction (random, 5 foci and one focus), with different values of the transmission rate (β) and range (r).

between the maximum and minimum percentage of S at 360 months was ranged from 0.22 to 5.4 with random introduction; this difference decreases as r increased, while different values of β produced similar results. With the two types of aggregate introduction and the lower values of r , the variability gradually increased in the first months and then almost stabilized, with the median being closer to the minimum. In the simulations with one initial focus located in an area at a greater distance than r of the rest of the individuals, the spread of the disease was practically null. The greatest variability was observed when the introduction was at one focus, with the largest difference with $r = 750$ m and $\beta = 0.015$ month⁻¹, where the percentage of S was 60% when the initial focus occurred toward the eastern end and 8% when the introduction was in the central area, leading to a greater disease spread (Fig. 6a, b). However, with the same r and β , this large difference was observed in earlier months, but after 360 months the difference decreased to 14.4%. The initial introduction at 5 foci had the highest variability in the final percentage of S with $r = 500$. Specifically, with $\beta = 0.03$ month⁻¹, the final percentage of S ranged from 79 to 38% (Fig. 6c, d). In these cases, the random introduction of the 5 foci resulted in two of them being very close together, even so, as shown in Fig. 6, the location of the foci and the distance between the individuals themselves was a determining factor in the progress of the disease. For both types of aggregate introduction and the higher value of r and β 0.015 or 0.03 month⁻¹, although differences were observed at the different times, the difference in the percentage of S at the end was zero.

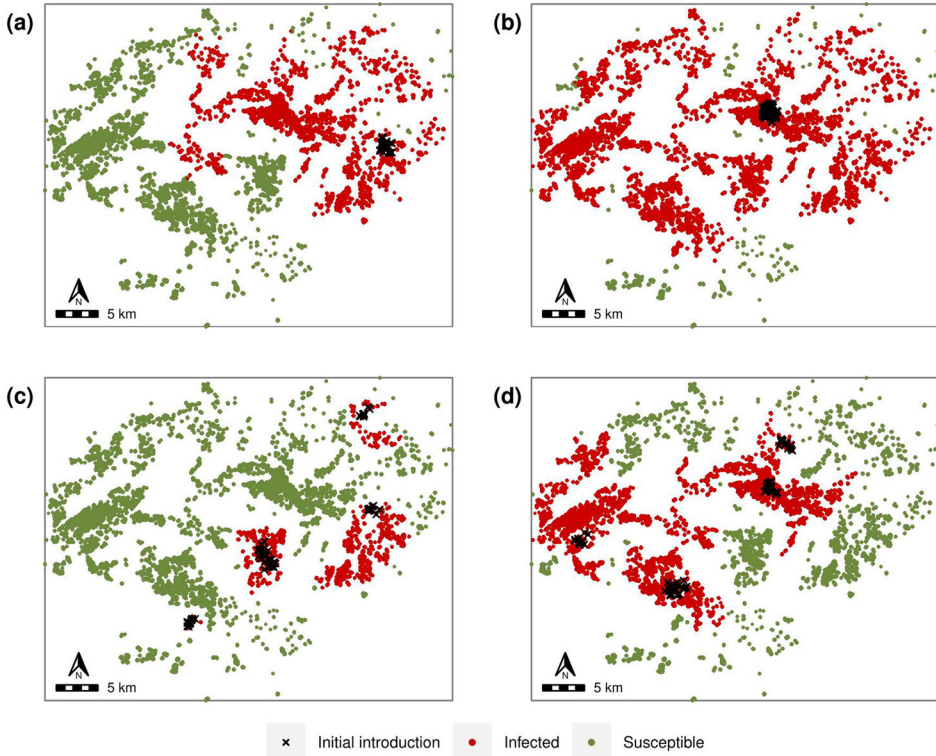


Figure 6. Simulations with which the greatest difference was obtained depending on the geographic location of initial introduction. **a, b** Minimum and maximum number of infected individuals after 360 months, respectively, with initial introduction at one focus, transmission rate $\beta = 0.015 \text{ month}^{-1}$ and range $r = 750 \text{ m}$. **c, d** Minimum and maximum infected individuals after 360 months, respectively, with initial introduction at 5 foci, transmission rate $\beta = 0.03 \text{ month}^{-1}$ and range $r = 500 \text{ m}$.

3.3.3. Variability due to Intrinsic Stochasticity

The intrinsic variability of the stochastic model was observed by running 10 times each combination of β , r , and the three types of initial introduction. For each type of introduction, the same configuration of the location of the initially infected individuals was fixed. Disease dynamics showed the same behavior as described above, with very little variability observed in the different simulations for each combination (Fig. 7). The only notable differences were those obtained in the last few months with the introduction at 5 foci, $r = 500 \text{ m}$, and $\beta = 0.03 \text{ month}^{-1}$, where the difference between the maximum and minimum percentage of S after 360 months was 19.3%. Despite this, the difference between the maximum and minimum of the final S percentage was between 0 and 5% with the introduction at 5 foci for the rest of the combinations. With random introduction, the difference between the maximum and minimum percentage of S at 360 months ranged from 0 to 1.2%, converging with the highest r values. With the introduction at one focus, the difference ranged from 0 to 9.6%, with the largest difference corresponding to the simulations performed with $r = 750 \text{ m}$ and $\beta = 0.005 \text{ month}^{-1}$.

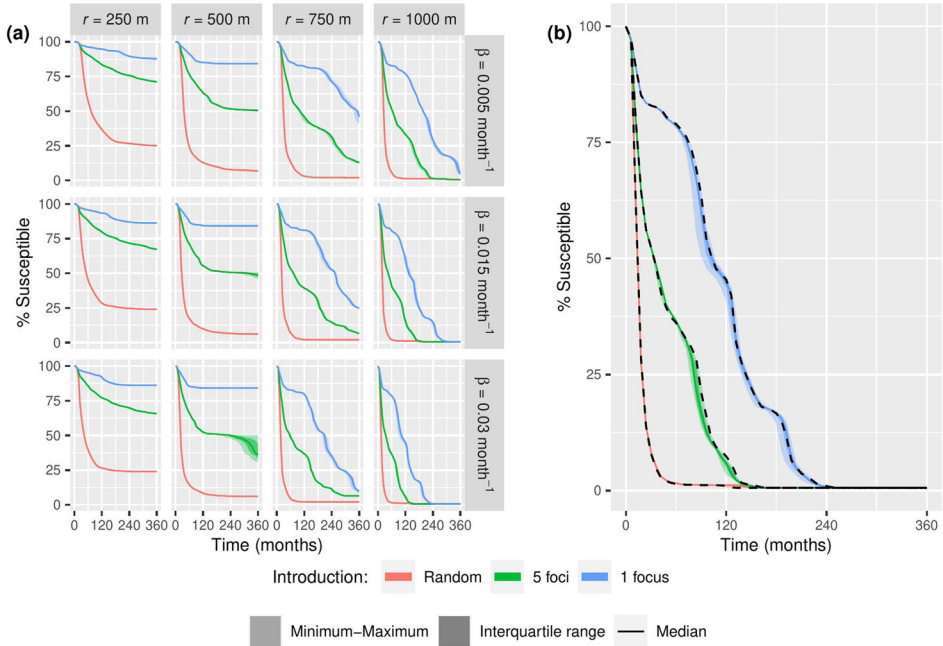


Figure 7. **a** Minimum, maximum, interquartile range and median of the percentage of susceptible individuals over time of the 10 simulations with each combination of the different values of the transmission rate (β), the range (r) of the Matern correlation function, and the type of initial introduction (random, 5 foci or one focus). **b** The dashed line corresponds to the simulations in which the probability of infection was calculated for each susceptible individual with $\beta = 0.03$ month⁻¹ and $r = 1000$ m, for each type of initial introduction (random, 5 foci and one focus), overlapping with the intervals of the 10 simulations obtained with the same values of the parameters β and r , where the probability of infection was calculated for each 1 ha cell.

In order to test whether calculating the probability of infection per 1 ha cell produced a bias in the results, the simulation of disease spread was performed by calculating the probability of infection for each individual S with $\beta = 0.03$ month⁻¹, $r = 1000$ m and each of the three types of introduction. The results obtained with these simulations did not differ from the approximation, with the percentage of S individuals within the interval obtained with the previous simulations (Fig. 7).

Depending on the values of the parameters r and β and the type of introduction, for simulations where the probability of infection was calculated as a function of the 1-ha cell grid, the computation time varied between 53 and 609 s, with the slower the disease progression, the longer the computation time. However, for the simulations where the probability of infection was calculated for each individual, the computational time varied between 12 and 36 min.

4. DISCUSSION

Epidemiological models are useful tools for the study of diseases, since they assist in the design and implementation of control strategies to avoid or minimize disease spread. Many factors can influence disease progress, including the environment, human interventions, and

means of spread, among others. Both the complexity of taking all these factors into account and the lack of suitable data make these models an approximation of reality. Therefore, model complexity and the factors to be taken into account are often determined by the specific objectives to be achieved. Compartmental models with differential equations are widely used in the study of disease epidemics, allowing an overall description of the progression of the disease at the population level. However, these models do not allow identification of relevant factors in the potential spread of the disease, such as disease status at the individual level and spatial heterogeneity.

The main objective of this work was to develop a tool for the study of disease spread at the individual plant level with heterogeneous spatial distribution. This model can be used as a basis for different applications including epidemiological surveillance, risk assessment, disease control, and impact studies. One of the main issues when dealing with models at the individual level is the high computational cost due to the large number of iterations. In this study, the algorithm was applied to a simplified and generalizable model based on three compartments (S , I_a and I_s) and their respective interactions, also integrating the spatial dependence. This model allowed us to simulate the spread of the disease under different conditions with feasible computational times. The spatial correlation was included by means of the Matérn correlation function, due to its flexibility and the interpretability of its parameters, which is widely used in diverse scientific disciplines (Guttorp and Gneiting 2006). Despite this, so far it has not been integrated in individual-based models in plant disease epidemiology.

When data are available that collect the number of individuals in each of the disease states at different times, it is feasible to estimate the parameters associated with the system. As detailed above, it can be approached from a deterministic perspective, however, opting for a stochastic or random scenario is more advisable due to the presence of measurement errors in the data and the influence of external factors that can impact the system. Two main classes of differential equations with uncertainty exist, stochastic differential equations (SDEs) and random differential equations (RDEs). In the case of SDEs, differential equations are influenced by an irregular stochastic process. A classic example of an SDE is a differential equation perturbed by a term dependent on a white noise variable, calculated as the derivative of the Wiener process or Brownian motion (Oksendal 2007; Gard 1988). To solve an SDE, the so-called Itô calculus can be employed, based on the application of the Itô lemma (Gard 1988). On the other hand, an RDE represents a natural generalization of the deterministic counterpart (Soong 1973; Strand 1970). In RDEs, the input parameters are considered random variables rather than fixed constants, respectively. Random differential equations constitute a natural generalization of their deterministic counterpart. Random effects are directly manifested in parameters, which are considered random variables or stochastic processes, i.e., parameters are assumed to have regular sample behavior described by standard probabilistic distributions. By incorporating a stochastic approach and considering the inherent uncertainty, both SDEs and RDEs offer valuable insights and flexibility when dealing with real-world models.

In our case, given the limited availability of data, we employed a compartmental model, which includes the spatial dependence between individuals using the Matérn function, to obtain a reference value for the transmission rate parameter. Subsequently, based on this and

information provided by the literature, an individual-based model is used for disease spread, where we compare how the parameters affect the solution of this spatiotemporal model. If data were available, and as previously indicated, we could have introduced uncertainty in the system of differential equations, considering the parameters as random variables. On the other hand, even if data from the origin of the introduction are not available, the simulation model used can provide a retrospective view.

The effects of the model parameters in disease spread were compared individually by simulating all possible combinations of the selected values. Due to the lack of suitable temporal data for inference on the β parameter, it was approximated to a reasonable value based on official surveillance data for *X. fastidiosa* in the study area. Due to the limitations of obtaining this parameter, other values around this parameter were also compared by the simulations. It was found that the r parameter of the Matérn correlation function as well as the type and location of the initial introduction were highly influential in the spread of the disease. For the highest values of r , almost the entire population was infected, regardless of β or the type of initial introduction. At lower r values, at certain times the spread of the disease was stopped, even though not all individuals were infected. The time of this interruption depended on the type of initial disease introduction and the value of r . The number of infected individuals increased with r , but the type of initial introduction also had a strong influence on the spread of the disease. When the initial introduction was random, with $r = 250\text{m}$ the incidence was even higher than when it was aggregated at a focus with $r = 500\text{m}$. These values of the range parameter can be associated with the distance of natural spread of the disease, i.e., by insect vectors, so it would be realistic to consider that if the insects have greater flight distance capacity they can infect more distant trees. Therefore, information on these could be used for the estimation of this parameter of the Matérn correlation function.

The variability of disease spread as a function of the type of initial introduction showed the importance of the location of this first introduction, particularly if it was aggregated. When the infection was initiated at an isolated location, the surrounding area without trees acted as a barrier to spread when r was less than the extent of this area where host plants are not present. On the contrary, in areas with higher population density, i.e., without large empty spaces, the infection rate was almost constant and depended on r . While a random introduction such as the one employed is unlikely to occur, aggregate introductions through the introduction of propagative material may more closely approximate reality.

Due to its simplicity, this individual-based model can be applied to the study of the spread of plant diseases caused by other pathogens by adapting the parameter values. Furthermore, with this basic proposal, the model could be extended with different disease states or dynamics adapted to each case study, such as the recovery or death of individuals, entry of new individuals into the population, elements of disease control, among others. However, it should be noted that more complexity in the model may imply a higher computational cost, as well as a larger number of parameters to estimate if data are available. Moreover, the complexity of the model would imply that it would be more case-specific and therefore less flexible.

In human disease epidemiology, the contact matrix is widely used to include the spatial relationship between individuals (Mahmood et al. 2021; Amaral et al. 2023). However, in

plant disease epidemiology, different spatial interactions have been associated with plant disease spread, depending on the host distribution, environmental conditions and the biology of the pathogen (Madden et al. 2007). The Matérn correlation function in this model allows the easy integration of diverse spatial relationships between individuals, thus making it capable of handling a variety of plant disease epidemics. This spread model can be used to establish areas at risk of disease, improve epidemiological surveillance, and optimize control strategies. Identifying the disease status of each individual allows the model to operate at a fine scale. With this, more realistic disease spread simulations are obtained and so the efficiency of control strategies can be better assessed.

ACKNOWLEDGEMENTS

The present work has received funding from grant E-RTA 2017-00004-C06-01 FEDER INIA AEI-MCIN and Organización Interprofesional del Aceite de Oliva Español and Horizon Europe program grant no. 101060593 (BeXyl, “Beyond Xylella, Integrated Management Strategies for Mitigating Xylella fastidiosa impact in Europe”). AL and DC are also grateful for grant PID2019-106341GB-I00 FEDER AEI-MCIN.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Declarations

Conflict of interest The authors report no conflict of interests.

Data Availability Data and code are available at <https://doi.org/10.5281/zenodo.10683923>. Results can be viewed at https://spatial-ibm.shinyapps.io/spread_results_app/.

[Received February 2023. Revised November 2023. Accepted January 2024.]

REFERENCES

- Akman O, Schaefer E (2015) An evolutionary computing approach for parameter estimation investigation of a model for cholera. *J Biol Dyn* 9(1):147–158
- Akman D, Akman O, Schaefer E (2018) Parameter estimation in ordinary differential equations modeling via particle swarm optimization. *J Appl Math* 2018:1–9
- Almeida RPP, Blua MJ, Lopes JRS, Purcell AH (2005) Vector transmission of *Xylella fastidiosa*: applying fundamental knowledge to generate disease management strategies. *Ann Entomol Soc Am* 98(6):775–786
- Amaral AVR, González JA, Moraga P (2023) Spatio-temporal modeling of infectious diseases by integrating compartment and point process models. *Stoch Env Res Risk Assess* 37(4):1519–1533
- Capaldi A, Behrend S, Berman B, Smith J, Wright J, Lloyd AL (2012) Parameter estimation and uncertainty quantification for an epidemic model. *Math Biosci Eng* 9(3):553–576
- Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2021) Shiny: web application framework for R. R package version 1.7.1

- Chen D, Moulin B, Wu J (2014) Analyzing and modeling spatial and temporal dynamics of infectious diseases. John Wiley & Sons, New Jersey
- Chowell G, Hyman JM, Bettencourt LMA, Castillo-Chávez C (2009) Mathematical and statistical estimation approaches in epidemiology. Springer, Dordrecht
- Cunniffe NJ, Koskella B, Metcalf CJE, Parnell S, Gottwald TR, Gilligan CA (2015) Thirteen challenges in modelling plant diseases. *Epidemics* 10:6–10
- Deardon R, Brooks SP, Grenfell BT, Keeling MJ, Tildesley MJ, Savill NJ, Shaw DJ, Woolhouse ME (2010) Inference for individual-models of infectious diseases in large populations. *Stat Sin* 20(1):239–261
- Diggle PJ, Ribeiro PJ, Christensen OF (2003) An introduction to model-based geostatistics. Springer, New York
- European Commission (2022) Emergency control measures by plant pest: *Xylella fastidiosa*. https://food.ec.europa.eu/plants/plant-health-and-biosecurity/legislation/control-measures/xylella-fastidiosa_en. Accessed 1 Oct 2022
- European Food Safety Authority (EFSA) (2019) Update of the scientific opinion on the risks to plant health posed by *Xylella fastidiosa* in the EU territory. *EFSA J* 17(5)
- European Food Safety Authority (EFSA) (2020) Update of the *Xylella* spp. host plant database—systematic literature search up to 30 June 2019. *EFSA J* 18(4)
- Gard T (1988) Introduction to stochastic differential equations. Marcel Dekker, New York
- Generalitat Valenciana (GVA) (2021) Plan de acción frente a *Xylella fastidiosa* en la Comunitat Valenciana. https://agroambient.gva.es/documents/163214705/163847802/20210719_Plan+accion_julio+2021.pdf/1ba03004-57fa-40f8-b79d-fc3214230057. Accessed 11 April 2022
- Gibson GJ (1997) Markov chain Monte Carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *J Roy Stat Soc Ser C (Appl Stat)* 46(2):215–233
- Gibson GJ, Kleczkowski A, Gilligan CA (2004) Bayesian analysis of botanical epidemics using stochastic compartmental models. *Proc Natl Acad Sci* 101(33):12120–12124
- Gilligan CA (2002) An epidemiological framework for disease management. *Adv Bot Res* 38:1–64
- Guttorp P, Gneiting T (2006) Studies in the history of probability and statistics XLIX on the Matérn correlation family. *Biometrika* 93(4):989–995
- Handcock MS, Wallis JR (1994) An approach to statistical spatial-temporal modeling of meteorological fields. *J Am Stat Assoc* 89(426):368–378
- Hyatt-Twynam SR, Parnell S, Stutt ROJH, Gottwald TR, Gilligan CA, Cunniffe NJ (2017) Risk-based management of invading plant disease. *New Phytol* 214(3):1317–1329
- Keeling MJ, Rohani P (2008) Modeling infectious diseases in humans and animals. Princeton University Press, Princeton
- Keeling MJ, Woolhouse ME, Shaw DJ, Matthews L, Chase-Topping M, Haydon DT, Cornell SJ, Kappey J, Wilesmith J, Grenfell BT (2001) Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape. *Science* 294(5543):813–817
- Kermack W, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proc R Soc Lond Ser A Contain Pap Math Phys Character* 115(772):700–721
- Kleczkowski A, Hoyle A, McMenemy P (2019) One model to rule them all? Modelling approaches across One-Health for human, animal and plant epidemics. *Philos Trans R Soc B Biol Sci* 374:20180255
- Lam SK, Pitrou A, Seibert S (2015) Numba: A LLVM-based Python JIT Compiler. In: Proceedings of the second workshop on the LLVM compiler infrastructure in HPC-LLVM'15. ACM Press, New York, pp 1–6
- Landa BB, Castillo AI, Giampetruzzi A, Kahn A, Román-Écija M, Velasco-Amo MP, Navas-Cortés JA, Marco-Noales E, Barbé S, Moralejo E, Coletta-Filho HD, Saldarelli P, Saponari M, Almeida RPP (2020) Emergence of a plant pathogen in Europe associated with multiple intercontinental introductions. *Appl Environ Microbiol* 86(3):e01521-19
- Lindgren F, Rue H, Lindström J (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J R Stat Soc Ser B (Stat Methodol)* 73(4):423–498

- Madden LV, Hughes G, Van Den Bosch F (2007) The study of plant disease epidemics. The American Phytopathological Society, St. Paul
- Mahmood M, Mateu J, Hernández-Orallo E (2021) Contextual contact tracing based on stochastic compartment modeling and spatial risk assessment. *Stoch Environ Res Risk Assess* 36:893–917
- Matérn B (1986) Spatial variation, 2nd edn. Springer, Berlin
- Matis J, Hartley H (1971) Stochastic compartmental analysis: model and least squares estimation from time series data. *Biometrics* 27(1):77–102
- Meentemeyer RK, Cunniffe NJ, Cook AR, Filipe JAN, Hunter RD, Rizzo DM, Gilligan CA, Cunniffe NJ, Cook AR, Filipe JAN, Hunter RD, Rizzo DM, Gilligan CA (2011) Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). *Ecosphere* 2(2):Article 17, 2:1–24
- Ministerio de Agricultura, Pesca y Alimentación (MAPA) (2021) Resumen de los datos contenidos en el SIGPAC para la campaña 2021
- Moralejo E, Gomila M, Montesinos M, Borràs D, Pascual A, Nieto A, Adrover F, Gost PA, Seguí G, Busquets A, Jurado-Rivera JA, Quetglas B, de Dios García J, Beidas O, Juan A, Velasco-Amo MP, Landa BB, Olmo D (2020) Phylogenetic inference enables reconstruction of a long-overlooked outbreak of almond leaf scorch disease (*Xylella fastidiosa*) in Europe. *Commun Biol* 3(560)
- Oksendal B (2007) Stochastic differential equations: an introduction with applications. Springer, Berlin
- Python Software Foundation (2021) The Python language reference—Python 3.8.12 documentation. Accessed 12 April 2022
- R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
- Saponari M, Boscia D, Nigro F, Martelli G (2013) Identification of DNA sequences related to *Xylella fastidiosa* in oleander, almond and olive trees exhibiting leaf scorch symptoms in Apulia (southern Italy). *J Plant Pathol* 95(3)
- Segura JMA, Cabetas MJR et al (2018) El almendro: diversidad de marcos de plantación. *Agricultura: Revista agropecuaria y ganadera* 1014:100–105
- Soong T (1973) Random differential equations in science and engineering. Academic Press, New York
- Stein ML (1999) Interpolation of spatial data. Springer, New York
- Strand J (1970) Random ordinary differential equations. *J Differ Equ* 7(3):538–553
- Wells JM, Raju BC, Hung HY, Weisburg WG, Mandelco-Paul L, Brenner DJ (1987) *Xylella fastidiosa* gen. nov., sp. nov: gram-negative, xylem-limited, fastidious plant bacteria related to *Xanthomonas* spp. *Int J Syst Evol Microbiol* 37(2):136–143
- White SM, Navas-Cortés JA, Bullock JM, Boscia D, Chapman DS (2020) Estimating the epidemiology of emerging *Xylella fastidiosa* outbreaks in olives. *Plant Pathol* 69(8):1403–1413

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.