Inferring invasion determinants with mechanistic models and multitype samples

M. Saubin[1], J. Coville[2], C. Xhaard[1,2,3], P. Frey[1], S. Soubeyrand[2], F. Halkett[1] and F. Fabre[4]

[1] Université de Lorraine, INRAE, IAM, F-54000 Nancy, France

[2] INRAE, BioSP, 84914 Avignon, France

[3] Université de Lorraine, INSERM CIC-P 1433, CHRU de Nancy, INSERM U1116, Nancy, France.

[4] INRAE, Bordeaux Sciences Agro, SAVE, F-33882 Villenave d'Ornon, France

Abstract

1- Dispersal, and in particular the frequency of long-distance dispersal (LDD) events, has strong implications for population dynamics, with possibly the acceleration of the colonisation front, and evolution, with possibly the conservation of genetic diversity along the colonised domain. However, accurately inferring LDD is challenging as it requires both large-scale data and a methodology that encompasses the redistribution of individuals in time and space.

2- Here, we propose a mechanistic-statistical framework to estimate dispersal of one-dimensional invasions. The mechanistic model takes into account population growth and grasps the diversity in dispersal processes by using either diffusion, leading to a reaction-diffusion (R.D.) formalism, or kernels, leading to an integro-differential (I.D.) formalism. The ID formalism considers different dispersal kernels (*e.g.* Gaussian, Exponential, and Exponential-power) differing in their frequency of LDD events. The statistical model relies on dedicated observation laws that describe two types of samples possibly gathered in space and time during the invasion (an overall survey and/or a refined examination of clumped samples) while taking into account the variability in both habitat suitability and occupancy perception.

3- We first check the identifiability of the parameters and the confidence in the selection of the dispersal process. We observed good identifiability for nearly all parameters (Correlation Coefficient > 0.95 between true and fitted values), except for occupancy perception (Correlation Coefficient $= 0.83 - 0.85$). The Exponential-Power (*i.e.* fat-tailed) kernel is the dispersal process most confidently identified. We then applied our framework to data describing an annual invasion of the poplar rust disease along the Durance River valley over nearly 200 km. This spatio-temporal survey consisted of 12 study sites examined at seven time points. We confidently estimated that the dispersal of poplar rust is best described by an Exponential-power kernel with a mean dispersal distance of 2.01 km and an exponent parameter of 0.24 characterising a fat-tailed kernel with frequent LDD events.

4- By considering the whole range of possible dispersal processes our method forms a robust inference method. It can be employed for a variety of organisms provided they are monitored in time and space along a one-dimension invasion.

Keywords

Long-distance dispersal, mechanistic-statistical model, Multiple data types,

1 Introduction

Dispersal is key in ecology and evolutionary biology (Clobert et al., 2004). From an applied point of view, the knowledge of dispersal is of prime interest for designing ecological-based management strategies in a wide diversity of contexts ranging from the conservation of endangered species (*e.g.*, Macdonald and Johnson, 2001) to the mitigation of emerging epidemics (Dybiec et al., 2009; Fabre et al., 2021). From a theoretical point of view, the pattern and strength of dispersal sharply impact eco-evolutionary dynamics (*i.e.* the reciprocal interactions between ecological and evolutionary processes) (Miller et al., 2020). Typically, besides the mean dispersal abilities, the finer features of dispersal have many implications for population dynamics (*e.g.* speed of invasion, metapopulation turnover; Soubeyrand et al., 2015; Kot et al., 1996), genetic structure (*e.g.* gene diversity, population differentiation; Edmonds et al., 2004; Fayard et al., 2009; Petit, 2011) and local adaptation (Gandon and Michalakis, 2002; Hallatschek and Fisher, 2014). Mathematically, the movement of dispersers (individuals, spores or propagules for example) can be described by a so-called location dispersal kernel (Nathan et al., 2012) that represents the statistical distribution of the locations of the propagules of interest after dispersal from a source point. Since the pioneer works of Mollison (1977), much more attention has been paid to the fatness of the tail of the dispersal kernel. When, at a relatively large distance, the shape of the tail of a kernel decreases less or equally slowly than an exponential distribution, the kernel is termed short-tailed (also referred to as thin-tailed). If it decreases more slowly, the kernel is termed long-tailed (also referred to as fat-tailed) (Klein et al., 2006). In this case, long-distance dispersal (LDD) events are more frequent than with an exponential kernel that shares the same mean dispersal distance. The frequency of LDD events has consequences on both population dynamics and genetic structure. Short-tailed dispersal kernels generate an invasion front of constant velocity, whereas long-tailed kernels cause an accelerating front of colonisation (Ferrandino, 1993; Kot et al., 1996; Clark et al., 2001; Mundt et al., 2009; Hallatschek and Fisher, 2014). Frequent LDD events can also cause a reshuffling of alleles along the colonisation gradient, which prevents the erosion of genetic diversity (Nichols and Hewitt, 1994; Petit, 2004; Fayard et al., 2009) or leads to patchy population structures (Ibrahim et al., 1996; Bialozyt et al., 2006).

Despite being a major issue in biology, properly characterising the dispersal kernels is a challenging task

for many species, especially when dispersing individuals are numerous, small (and thus difficult to track) and move far away (Nathan, 2001). This is typically the case of plant pathogens for which dispersal kernels have rarely been assessed (Fabre et al., 2021) despite their major impact on food security and ecosystems (Strange and Scott, 2005; Savary et al., 2019). Three approaches can be distinguished to infer dispersal (Nathan et al., 2003; Rieux et al., 2014). The first class of methods uses the indirect effect of dispersal on the pattern of genetic differentiation among populations to estimate dispersal parameters. Most of them only estimate an average dispersal distance or a diffusion rate (Broquet and Petit, 2009; Roques et al., 2016). The second class enables to infer the entire shape of the dispersal kernel by directly observing dispersal with dedicated experiments tracking the movement of tagged individuals, thanks to particular morphological structures or genetic markers. They often rely on heavy experimental procedures for data acquisition, requiring for example to control for the source strength and location (Soubeyrand et al., 2007; Rieux et al., 2014). The third class of methods make use of specific mathematical modelling to infer dispersal from observations of ongoing dynamics (Soubeyrand et al., 2009a; Rieux et al., 2013; Bousset et al., 2015; Grosdidier et al., 2018). If phenomenological models can hardly disentangle the observed data from the dispersal process *per se*, mechanistic-statistical models do so and enable a proper inference of dispersal using spatio-temporal datasets (Wikle, 2003a; Soubeyrand et al., 2009a; Roques et al., 2011; Soubeyrand and Roques, 2014; Hefley et al., 2017; Nembot Fomba et al., 2021). These models allow for the parsimonious representation of both growth and dispersal processes in heterogenous environments (Papaïx et al., 2022). They require detailed knowledge of the biology of the species of interest to properly model the invasion process. They combine a mechanistic model describing the invasion process and a probabilistic model describing the observation process while enabling a proper inference using spatio-temporal data.

Classically, the dynamics of large populations are well described by deterministic differential equations. Following the seminal work of Fisher on the spread of a mutant gene in a given population (Fisher, 1937), invasions have often been modelled through reaction-diffusion equations (Murray, 2002; Okubo and Levin, 2002; Shigesada and Kawasaki, 1997). In this setting, individuals are assumed to move randomly in their environment. Their trajectories are modelled using a Brownian motion or a more general stochastic diffusion process. Whereas reaction-diffusion equations have been classically used to describe biological invasions, their

incorporation into mechanistic-statistical approaches to estimate parameters of interest from spatio-temporal data essentially dates back to the early 2000s. Examples include the study of the invasion processes of the House Finch bird over the eastern United States (Wikle, 2003a), the pine processionary moth in France (Roques et al., 2011; Soubeyrand and Roques, 2014), the wolf in Eastern France (Louvrier et al., 2020) and the spread of the black pod disease in a cocoa plot (Nembot Fomba et al., 2021). By contrast to reaction-diffusion equations, integro-differential equations rely on dispersal kernels, individuals being redistributed according to the considered kernel (Fife, 1996; Hutson et al., 2003; Kolmogorov et al., 1937). This approach allows to consider a large variety of dispersal functions, typically with either a short or a long tail (*i.e.* putative LDD events). As such it is more likely to model accurately the true organisms dispersal process. However, as far as we know, integro-differential equations have rarely been embedded into mechanistic-statistical approaches to infer dispersal processes in ecology (but see Szymańska et al., 2021 for a recently proposed application of a non-local model to cell proliferation).

Data acquisition is another challenge faced by biologists in the field, all the more that data confined to relatively small spatial scales can blur the precise estimates of the shape of the kernels tail (Ferrandino, 1996; Kuparinen et al., 2007; Rieux et al., 2014). To gather as much information as possible, it is mandatory to collect data over a wide range of putative population sizes (from absence to near saturation) along the region of interest. In turn, the practical implementation of such a sampling scheme may be difficult to grasp from a single data type as one faces a trade-off in data acquisition between a large-scale monitoring (*e.g.* the screening of a large number of putative habitat units to test for presence/absence, thereafter raw samples) and, when present, a small-scale observation (*e.g.* the focusing on a reduced number of habitat units to estimate more precisely local population sizes and individual aggregation, thereafter refined samples). Sharing the sampling effort between raw and refined samples to browse through the propagation front may improve the inference of spatial ecological processes (Gotway and Young, 2002). Furthermore, the probabilistic model describing the observation process in the mechanistic-statistical approach can handle such multiple datasets with varying supports (Wikle, 2003b). However, inference based on multi-type data remains a challenging statistical issue as the observation variables describing each data type follow different distribution laws (Chagneau et al.,

2011) and can be correlated or, more generally, dependent because they are governed by the same underlying dynamics (Bourgeois et al., 2012; Georgescu et al., 2014; Soubeyrand et al., 2018). This requires careful definition of the conditional links between the observed variables and the model parameters (the so-called observation laws) in order to identify and examine complementarity and possible redundancy between the data types.

In this article, we aim to provide a sound and unified inferential framework to estimate dispersal from ecological invasion data using both reaction-diffusion and integro-differential equations. For the sake of simplicity we focus on one-dimensional invasions, but our approach can be extended to 2D-analyses as well. We first define the two classes of mechanistic invasion models, establish the observation laws corresponding to raw and refined samplings, and propose a maximum-likelihood method to estimate their parameters within the same inferential framework. Then, to confirm that each model parameter can indeed be efficiently estimated given the amount of data at hand (see Soubeyrand and Roques, 2014), we perform a simulation study to check model-parameters identifiability given a proposed experimental design. We also aim to assess the confidence level in the choice of the dispersal function as derived by model selection. Last, the inferential framework is applied to original ecological data describing the annual invasion of a tree pathogen (*Melampsora larici-populina*, a fungus species responsible for the poplar rust disease) along the riparian stands of wild poplars bordering the Durance River valley in the French Alps (Xhaard et al., 2012).

2 Modelling one-dimensional invasion and observation processes

2.1 A class of deterministic and mechanistic invasion models

In this study, we model the dynamics of a population density $u(t, x)$ at any time t and point x during an invasion using two types of spatially heterogeneous deterministic models allowing to represent a wide range of dispersal processes. Specifically, we considered a reaction-diffusion model (R.D.) and an integro-differential model (I.D.):

$$\text{R.D.}\begin{cases}\partial_t u(t,x) = D\partial_{xx}u(t,x) + r(x)u(t,x)\left(1-\frac{u(t,x)}{K}\right),\\ u(0,x) = u_0(x),\end{cases} \qquad \text{I.D.}\begin{cases}\partial_t u(t,x) = \int_{-R}^{R} J(x-y)[u(t,y)-u(t,x)]\,dy + r(x)u(t,x)\left(1-\frac{u(t,x)}{K}\right),\\ u(0,x) = u_0(x).\end{cases}$$

where t varies in $[0,T]$ (*i.e.* the study period) and x varies in $[-R,R]$ (*i.e.* the study domain). Both equations exhibit the same structure composed of a diffusion/dispersal component and a reaction component. The reaction component, $r(x)u(t,x)\left(1-\frac{u(t,x)}{K}\right)$ in both equations, is parameterised by a spatial growth rate $r(x)$ that takes into account macro-scale variations of the factors regulating the population density and K the carrying capacity of the environment. It models population growth. The diffusion/dispersal component models population movements either by a diffusion process ($D\partial_{xx}u$ in R.D.) parameterised by the diffusion coefficient D or by a dispersal kernel (J in I.D.) depending on a specific set of parameters (see below). To cover a large spectrum of possible dispersal processes, we use the following parametric form for the kernel J:

$$J := \frac{\tau}{2\alpha\Gamma\left(\frac{1}{\tau}\right)}e^{-\left|\frac{z}{\alpha}\right|^{\tau}}$$

with mean dispersal distance $\lambda := \alpha\frac{\Gamma\left(\frac{2}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right)}$. Varying the value of τ leads to kernels classically used in dispersal studies. Specifically, J can be a Gaussian kernel ($\tau = 2, \lambda = \alpha/\sqrt{\pi}$), an exponential kernel ($\tau = 1, \lambda = \alpha$) or a fat-tail kernel ($\tau < 1, \lambda = \alpha\Gamma\left(\frac{2}{\tau}\right)/\Gamma\left(\frac{1}{\tau}\right)$). Explicit formulas for the solution $u(t,x)$ of these reaction-diffusion/dispersal equations being out of reach, we compute a numerical approximation u_{num} of u which serves as a surrogate for the real solution. Details of the numerical scheme used to compute u_{num} can be found in Appendix A.

2.2 A conditional stochastic model to handle micro-scale fluctuations

Among the factors driving population dynamics, some are structured at large spatial scales (macro-scale) and others at local scales (micro-scale). Both scales may be considered when studying biological invasions. In the model just introduced, the term $r(x)$ describes factors impacting population growth rate at the macro-scale along the whole spatial domain considered. Accordingly, the function $u(t,x)$ is a mean-field approximation of the true population density at macro-scale. Furthermore, the population density can fluctuate due to micro-scale variations of other factors regulating population densities (*e.g.* because of variations in the micro-

climate and the host susceptibility). To take into account such local fluctuations, the deterministic invasion model is combined with a conditional probability distribution of the local population size in a habitat unit. The habitat unit is typically a small fraction of the total habitat. This distribution is conditional on $u(t,x)$, the macro-scale population density, and depends on the (unobserved) suitability of the habitat unit. The following paragraphs detail this dependence.

Consider a habitat unit i whose centroid is located at x_i, and suppose that the habitat unit is small enough to reasonably assume that $u(t,x) = u(t,x_i)$ for all location x in the habitat unit. Let $N_i(t)$ denote the number of individuals in i at time t. The conditional distribution of $N_i(t)$ is modelled by a Poisson distribution:

$$N_i(t) \mid u(t,x_i), R_i(t) \sim \text{Poisson}(u(t,x_i)R_i(t)), \tag{1}$$

where $R_i(t)$ is the intrinsic propensity of the habitat unit i to be occupied by individuals of the population at time t. Thereafter, $R_i(t)$ is called habitat suitability and takes into account factors like the exposure and the favorability of habitat unit i. The suitability of habitat unit i is a random effect (unobserved variable) and is assumed to follow a Gamma distribution with shape parameter σ^{-2} and scale parameter σ^2:

$$R_i(t) \sim \text{Gamma}(\sigma^{-2}, \sigma^2). \tag{2}$$

This parameterisation implies that the mean and variance of $R_i(t)$ are 1 and σ^2, respectively; that the conditional mean and variance of $N_i(t)$ given $u(t,x_i)$ are $u(t,x_i)$ and $u(t,x_i) + u(t,x_i)^2\sigma^2$, respectively; and that its conditional distribution is:

$$N_i(t) \mid u(t,x_i) \sim \text{Negative-Binomial}\left(\sigma^{-2}, \frac{u(t,x_i)\sigma^2}{1+u(t,x_i)\sigma^2}\right). \tag{3}$$

2.3 Multi-type sampling and models for the observation processes

During an invasion, the population density may range from zero (beyond the front) to the maximum carrying capacity of the habitat. To optimise the sampling effort, it may be relevant to carry out different sampling procedures depending on the population density at the sampling sites. In this article, we consider a two-stage

sampling made of one raw sampling which is systematic and one optional refined sampling adapted to our case study, the downstream spread of a fungal pathogen along a river. We consider that the habitat unit is a leaf. The fungal population is monitored in sampling sites $s \in \{1, \ldots, S\}$ and at sampling times $t \in \{t_1, \ldots, t_K\}$. Sampling sites are assumed to be small with respect to the study region and the duration for collecting one sample is assumed to be short with respect to the study period. Thus, the (macro-scale) density of the population at sampling time t in sampling site s is constant and equal to $u(t, z_s)$ where z_s is the centroid of the sampling site s and u satisfies the R.D. or I.D. equation. Any sampling site s is assumed to contain a large number of leaves which are, as a consequence of the assumptions made above, all associated with the same population density function: $u(t, x_i) = u(t, z_s)$ for all leaves i within sampling site s.

The raw sampling is focused on trees, considered as a group of independent leaves. In each sampling site s and at each sampling time t, a number B_{st} of trees is monitored for the presence of infection. We count the number of infected trees Y_{st} among the total number B_{st} of observed trees. In the simulations and the case study tackled below, the random variables Y_{st} given $u(t, x_s)$ are independent and distributed under the conditional probability distribution f_{st}^{raw} described in Appendix B.2. The distribution f_{st}^{raw} is a Binomial distribution. Its success probability depends on two parameters. It depends first on the variability of the biological process, through the variance parameter σ^2 of habitat suitabilities. It also depends on the variability of the observation process, through a parameter γ describing how the probabilities of leaf infection perceived by the person in charge of the sampling differ between trees from true probabilities (as informed by the mechanistic model). Such differences may be due for example to the specific configuration of the canopy of each tree or to particular lighting conditions.

The refined sampling is focused on twigs, considered as a group of connected leaves. Nearby leaves often encounter the same environmental conditions and, therefore, are characterised by similar habitat suitabilities represented by $R_i(t)$; see Equations (1–2). This spatial dependence was taken into account by assuming that the leaves of the same twig (considered as a small group of spatially connected leaves) share the same leaf suitability. Accordingly, suitabilities are considered as shared random effects. The refined sampling is performed depending on disease prevalence and available time. In site s at time t, G_{st} twigs are collected. For each twig g, the total number of leaves M_{stg} and the number of infected leaves Y_{stg} are counted. In

the simulations and the case study tackled below, the random variables Y_{stg} given $u(t, x_s)$ are independent and distributed under conditional probability distributions denoted by f_{st}^{ref} described in Appendix B.3. The distribution f_{st}^{ref} is a new mixture distribution (called Gamma-Binomial distribution) obtained using Equations (1–2) and taking into account the spatial dependence and the variance parameter of unobserved suitabilities (see Appendix B.3).

This sampling scheme and its vocabulary (leaves, twigs and trees) is specifically adapted to our case study for the sake of clarity. However, a wide variety of multi-type sampling strategies can be defined and implemented in the model, as long as it follows a two-stage sampling as presented in Figure 1.

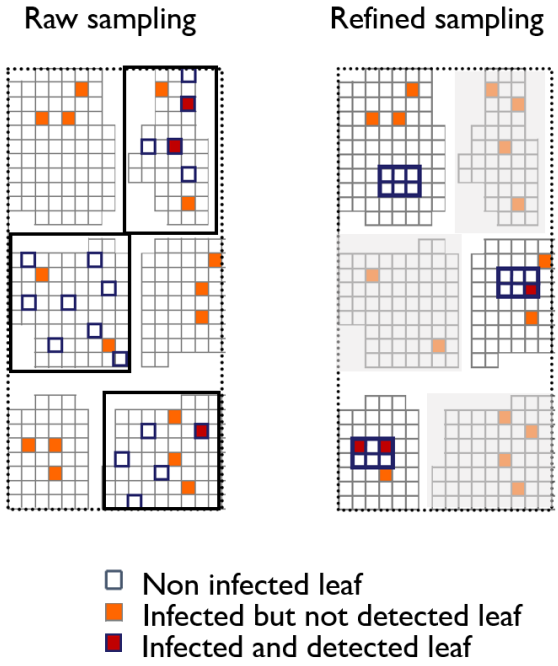

Figure 1: Two-stage sampling on a sampling site, with one systematic raw sampling (on the left) and one optional refined sampling (on the right). Each square represent a leaf, which can be non infected, infected but not detected, or infected and detected. Each group of spatially grouped leaves represent a tree. Each tree already observed during the raw sampling are not available (and thus represented in grey) for the refined sampling, where connected leaves in twigs are observed.

2.4 Coupling the mechanistic and observation models

The submodels of the population dynamics and the observation processes described above can be coupled to obtain a mechanistic-statistical model (also called physical-statistical model; Berliner, 2003; Soubeyrand

et al., 2009b) representing the data and depending on dynamical parameters, namely the growth and dispersal parameters. The likelihood of this mechanistic-statistical model can be written:

$$L(\theta) = \prod_{s=1}^{S} \prod_{t=t_1}^{t_K} \left\{ f_{st}^{\text{raw}}(Y_{st}) \left(\prod_{g=1}^{G_{st}} f_{st}^{\text{ref}}(Y_{stg}) \right)^{\mathbf{1}(\mathbf{Y_{st}} > \bar{\mathbf{y}})} \right\}, \tag{4}$$

where expressions of f_{st}^{raw} and f_{st}^{ref} adapted to the case study tackled below are given by Equations (S10) and (S14) in Appendix B. The power $\mathbf{1}(\mathbf{Y_{st}} > \bar{\mathbf{y}})$ (that is equal to 1 if $Y_{st} > \bar{y}$, 0 otherwise) implies that the product $\prod_{g=1}^{G_{st}} f_{st}^{\text{ref}}(Y_{stg})$ only appears if the refined sampling is carried out in site s. Moreover, such a product expression for the likelihood is achieved by assuming that leaves in the raw sampling and those in the refined sampling are not sampled from the same trees. If this does not hold, then an asymptotic assumption like the one in Appendix B.2 can be made to obtain Equation (4), or the dependence of the unobserved suitabilities must be taken into account and another likelihood expression must be derived.

3 Parameter estimation and model selection

We performed simulations to check the practical identifiability of several scenarios of biological invasions. Invasion scenarios represent a wide range of possible states of nature regarding the dispersal process, the environmental heterogeneity at macro-scale and the intensity of local fluctuations at micro-scale. Even though the simulations are designed to cope with the structure of our real data set (Appendix D), the results enable some generic insights to be gained. Specifically, we considered six sampling dates evenly distributed in time, and 12 samplings sites evenly distributed within the 1D spatial domain. For each pair *(date, site)*, we simulated the raw sampling of 100 trees and the refined sampling of 20 twigs. For the fifth sampling date, the raw sampling was densified with 45 sampling sites instead of 12.

The simulation study explored four hypotheses for the dispersal process: three I.D. hypotheses with kernels J_{Exp}, J_{Gauss} and J_{ExpP} and the R.D. hypothesis. Hypotheses J_{Exp} and J_{Gauss} state that individuals dispersed according to Exponential and Gaussian kernels, respectively, with parameter $\theta_J = (\lambda)$. Hypothesis J_{ExpP} states that individuals dispersed according to a fat-tail Exponential-power kernel with parameters $\theta_J = (\lambda, \tau)$ and $\tau < 1$. Finally, hypothesis R.D. states that individual dispersal is a diffusion process parameterised by

$\theta_J = (\lambda)$. The parameter λ represents the mean distance travelled whatever the dispersal hypothesis considered. Moreover, macro-scale environmental heterogeneity was accounted for in the simulations by varying the intrinsic growth rate of the pathogen population (r) in space. Specifically, along the one-dimensional domain, we considered two values of r, namely a downstream value r_{dw} and an upstream value r_{up}, parameterised by $\theta_r = (r_{\text{dw}}, \omega)$ such that $r_{\text{up}} = r_{\text{dw}} e^{\omega}$. Finally, micro-scale heterogeneity was accounted for in the simulations by varying the parameter of leaf suitability σ^2 and tree perception γ. Thereafter, $\theta = (\theta_r, \theta_J, \gamma, \sigma^2)$ denotes the vector of model parameters.

3.1 Accurate inference of model parameters

To assess the estimation method and check if real data that were collected are informative enough to efficiently estimate the parameters of the models (the so-called practical identifiability), we proceeded in three steps for each dispersal hypothesis: (i) a set of parameter values $\theta = (\theta_r, \theta_J, \gamma, \sigma^2)$ is randomly drawn from a distribution that encompasses a large diversity of realistic invasions, (ii) a data set with a structure similar to our real sampling is simulated given θ and (iii) θ is estimated using the maximum-likelihood method applied to the simulated data set. These steps were repeated $n = 100$ times. Details on the simulation procedure, the conditions used to generate realistic invasions and on the estimation algorithm are provided in Appendix D.1.

Practical identifiability was tested by means of correlation coefficients between the true and estimated parameter values for the four models corresponding to the four dispersal processes (see Table 1, Figures S2, S3, S4, S5).

All the parameters defining the macro-scale mechanistic invasion model (r_{dw}, ω, λ) display very good practical identifiability whatever the model, with correlation coefficients above 0.93. In the case of the Exponential-power dispersal kernel, the additional parameter representing the tail of the distribution (τ) also displays a very good practical identifiability with a correlation coefficient of 0.95. The parameter defining the micro-scale fluctuations, σ^2, lead to particularly high correlation coefficients (0.99 for all the models). The identifiability for the perception parameter γ related to the observation process is somewhat lower (from 0.83 to 0.85).

Table 1: Model practical identifiability. Numbers indicate the coefficient of correlation between the true and estimated parameter values for the four models corresponding to the four dispersal processes (J_{Exp}, J_{Gauss}, J_{ExpP} and R.D.) from 100 replicates. High correlation between true and estimated parameters indicates a good practical identifiability. The standard deviations of the coefficients of correlation, estimated with a bootstrapping method, are indicated in brackets. Correlation coefficients and standard deviations are given for natural scale for parameter ω, and logarithm scales for parameters r_{dw}, γ, λ, τ, and σ^2.

Parameter	Description	J_{Exp}	J_{Gauss}	J_{ExpP}	R.D.
r_{dw}	Growth rate downstream	$0.99(1.10^{-3})$	$0.99(1.10^{-3})$	$0.99(2.10^{-3})$	$0.93(6.10^{-2})$
ω	Growth rate modulator	$0.99(< 10^{-3})$	$0.99(< 10^{-3})$	$0.99(1.10^{-3})$	$0.99(1.10^{-3})$
λ	Mean dispersal distance	$0.99(5.10^{-3})$	$0.98(8.10^{-3})$	$0.99(1.10^{-3})$	$0.95(2.10^{-2})$
τ	Kernel exponent	NA	NA	$0.95(1.10^{-2})$	NA
γ	Tree perception	$0.85(4.10^{-2})$	$0.83(4.10^{-2})$	$0.83(5.10^{-2})$	$0.84(3.10^{-2})$
σ^2	Variance in leaf suitability	$0.99(1.10^{-3})$	$0.99(< 10^{-3})$	$0.99(< 10^{-3})$	$0.99(< 10^{-3})$

3.2 Confidence in the selection of the dispersal process

Numerical simulations were next designed to test whether model selection could disentangle the true dispersal process (*i.e.* the dispersal hypothesis used to simulate the data set) from alternative dispersal processes. We proceeded as indicated for model practical identifiability, except that we fitted to each data set the true model (as previously) but also the three other models corresponding to alternative dispersal hypotheses. We made $n = 50$ replicates for each of the 16 combinations (four models used for simulation times four models used for estimation). Models were ranked using AIC (Akaike Information Criteria, see Appendix D.2) to select the model best supported by the data set. Confidence level in model selection was assessed by the proportion of cases where the true model was selected according to AIC (Table 2). Recall that, as we only consider $\tau < 1$ in our numerical simulations (Appendix D.1), the dispersal kernel J_{ExpP} is always a fat-tail kernel.

Table 2: Efficiency of model selection using Akaike information criterion (AIC). The four first columns indicate the proportion of cases, among 50 replicates, where each tested model was selected using AIC, given that data sets were generated under a particular model (*i.e.* true model). Column $d\text{AIC}_{\text{true}}$ (*resp.* $d\text{AIC}_{\text{wrong}}$) indicates the mean difference between the AIC of the model selected when the model selected is the true one (*resp.* when the model selected is not the true model) and the second best model (*resp.* being the true model or not).

	Selected Model					
True Model	J_{Exp}	J_{Gauss}	J_{ExpP}	R.D.	$d\text{AIC}_{\text{true}}$	$d\text{AIC}_{\text{wrong}}$
J_{Exp}	**0.62**	0.22	0.06	0.10	0.84	0.74
J_{Gauss}	0.34	**0.26**	0.00	0.40	1.08	0.55
J_{ExpP}	0.20	0.04	**0.70**	0.06	**89.62**	0.38
R.D.	0.18	0.24	0.00	**0.58**	0.71	0.23

The model selection procedure is efficient for the dispersal hypotheses Exponential-power J_{ExpP}, Exponential J_{Exp}, and reaction-diffusion R.D., with 70%, 62% and 58% of correct kernel selection, respectively (Table 2). When the fat-tail Exponential-power kernel is not correctly identified, it is mostly mistaken with the Exponential one (for 20% of the simulations). In line with this, the probability of correctly selecting the kernel J_{ExpP} decreases when the parameter τ increases towards 1, the value for which the Exponential-power kernel coincides with the Exponential kernel (Figure 2). Importantly, when the Exponential-power kernel is correctly selected, we observe a large difference between its AIC and the AIC of the second best model (89.62 points on average). Conversely, when the invasion process is simulated under J_{ExpP}, but another kernel is selected, we observe a very small AIC difference (0.38 point on average). Model selection does not allow to correctly select the Gaussian kernel J_{Gauss} (Table 2). Indeed, with only 26% of correct model selection, this kernel is not better identified than with a random draw of one of the four models, which would lead to 25% of correct estimations. Its correct identification is greatly improved by densifying the sampling scheme (Appendix D.5, Table S3). Finally, note that when the invasion process is simulated under model R.D. or J_{Gauss}, a short-tail kernel is always selected and, thus, never confounded with the fat-tail kernel J_{ExpP}.

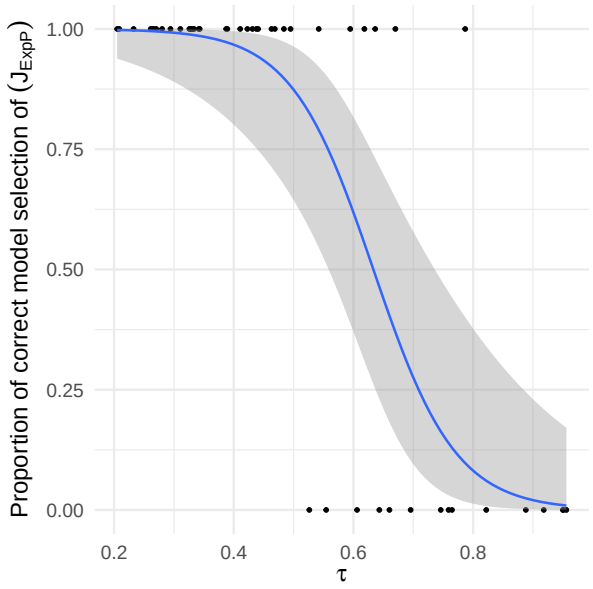

Figure 2: Logistic regression of the proportion of correct model selection of dispersal J_{ExpP} as a function of τ. Dots represent the values of τ used for the 50 replicates of simulated dispersal model J_{ExpP}, and the estimated dispersal model (1 for a correct model selection of J_{ExpP} and 0 for a wrong model selection). The blue line corresponds to the predicted value of the proportion of correct model selection J_{ExpP} as a function of τ, and the grey area corresponds to the confidence envelope at 95%.

4 Case study: Invasion of poplar rust along the Durance River valley

4.1 Study site

We applied our approach to infer the dispersal of the plant pathogen fungus *Melampsora larici-populina*, responsible for the poplar rust disease, from the monitoring of an epidemic invading the Durance River valley. Embanked in the French Alps, the Durance River valley constitutes a one-dimension ecological corridor which channels annual epidemics of the poplar rust pathogen *M. larici-populina* (Xhaard et al., 2012). Each year the fungus has to reproduce on larches (*Larix decidua*) that are located in the upstream part of the valley only. This constitutes the starting point of the annual epidemics. Then the fungus switches to poplar leaves and performs several rounds of infection until leaf-fall. Each infected leaf produces thousands of spores that are wind-dispersed. In our case study, $u(t, x_s)$ is the density of fungi infection at time t at point x on a poplar leaf. Each leaf has a carrying capacity of 750 lesions, Appendix E).

All along the valley, the Durance River is bordered by a nearly continuous riparian forest of wild poplars (*Populus nigra*). The annual epidemic on poplars thus spreads downstream through the riparian stands, mimicking a one-dimension biological invasion (Xhaard et al., 2012). A previous genetic study showed that the epidemic was indeed initiated in an upstream location where poplars and larchs coexist (Prelles) and progresses along the valley (Becheler et al., 2016). In fall, the corridor is cleared for disease after leaf-fall. At 62 km downstream of the starting point of the epidemics, the Serre-Ponçon dam represents a shift point in the valley topology, with a steed-sided valley upstream and a larger riparian zone downstream. This delimitation led us to consider 2 values of growth rates r along the one-dimensional domain: r_{up} and r_{dw} (see Appendix D for details).

4.2 Monitoring of an annual epidemic wave

In 2008, rust incidence was monitored every three weeks from July to November at 12 sites evenly distributed along the valley (Figure 3). Sites were inspected during seven rounds of surveys. For a unique date (Oct. 22), the raw sampling was densified with 45 sites monitored instead of 12. We focused on young poplar trees (up

311 to 2m high) growing on the stands by the riverside.

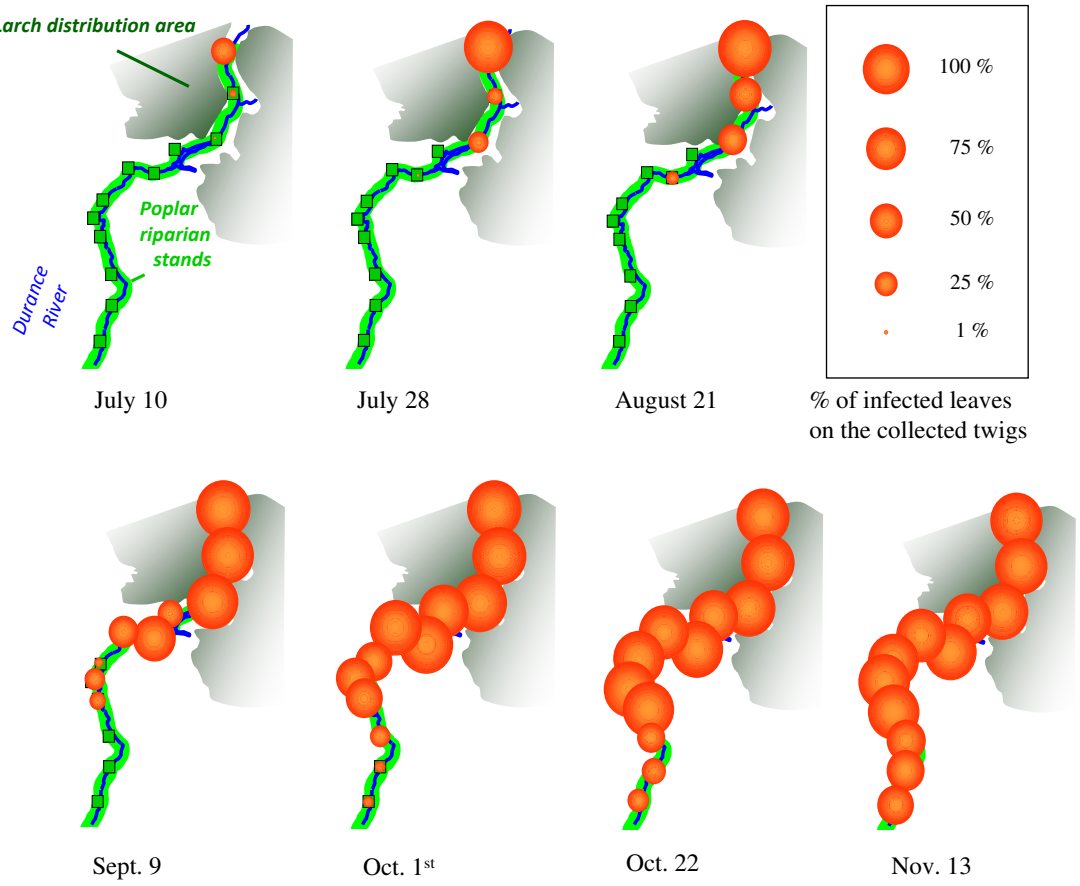

Figure 3: Poplar rust epidemic wave along the Durance River valley in 2008. The larch distribution area is represented in dark green, wild poplar riparian stands in pale green. The 12 study sites are represented by the green squares. Orange dots describe the evolution of the poplar rust epidemic through time (7 rounds of disease notation) and space (12 studied sites). Dot size is proportional to rust disease incidence assessed from the refined sampling.

312 Two monitorings were conducted, corresponding to the raw and refined sampling, as described in previous
313 sections. For the raw sampling, we prospected each site at each date to search for rust disease by inspecting
314 randomly distributed poplar trees (different trees at different dates for a given site). Depending on rust
315 incidence and poplar tree accessibility, 40 to 150 trees (mean 74) were checked for disease. Each tree was
316 inspected through a global scan of the leaves on different twigs until at least one infected leaf was found or
317 after 30 s of inspection. The tree was denoted infected or healthy, respectively. This survey method amounts
318 to minutely inspecting 10 leaves per tree, *i.e.* with the same efficiency of disease detection as through the

refined sampling (see details of the statistical procedure in Appendix C). The global scan procedure of the trees leads to equivalently surveying fewer and fewer leaves as the epidemic progresses. Optionally, when at least one tree was infected, and depending on available time, we carried out a refined sampling to collect more information on the variance in disease susceptibility (*i.e.* habitat suitability) among the sampling domain. The refined sampling consisted in randomly sampling 20 twigs on different trees and recording, for each, the total number of leaves and the number of infected leaves.

4.3 Dispersal and demographic processes ruling the epidemic wave

The inferential framework developed and tested in previous section was used to fit and compare the four models considered to the real dataset gathered in the Durance River valley and to derive confidence intervals of the parameters of interest (Appendix D.3). Model selection was used to decipher which dispersal process was best supported by the data set, for five initial conditions. The large AIC difference in favour of hypothesis J_{ExpP} indicates that poplar rust propagules assuredly disperse according to an exponential-power dispersal kernel along the Durance River valley (Table 3). Note that for all kernels, the five initial conditions lead to similar estimations. Under the R.D. hypothesis, however, initial conditions can lead to different estimations because of local optima, but all AIC resulting from the R.D. hypothesis are higher than AIC resulting from the three dispersal kernels.

Table 3: Model selection for the epidemic of poplar rust along the Durance River valley. The Akaike information criteria are indicated for each model fitted to the real data set. The model best supported by the data is indicated in bold. $\mathrm{AIC}_{\mathrm{median}}$ and $\mathrm{AIC}_{\mathrm{sd}}$ represent the median and standard deviation among the AIC obtained from five initial conditions.

Dispersal	$\mathrm{AIC}_{\mathrm{median}}$	$\mathrm{AIC}_{\mathrm{sd}}$
J_{Exp}	5476	0.68
J_{Gauss}	5510	1.03
$\boldsymbol{J_{\mathrm{ExpP}}}$	**5179**	**1.32**
R.D.	6303	655.60

The estimation of the parameters for the best model along with their confidence intervals are summarised in Table 4. The parameters of the Exponential-power kernel are of prime interest in our study. They firstly indicate that the mean distance travelled by rust spores is estimated at 2.01 km. Moreover, the mean exponent parameter τ of the Exponential-power kernel is 0.24. This value, much lower than 1, suggests substantial long-distance dispersal events. We also estimated the growth rates of the poplar rust epidemics along the

Durance River valley. From upstream to downstream, their mean estimates are 0.084 and 0.020, respectively. The estimate of the parameter of the observation model, γ, is 5.21. This parameter represents how perceived probabilities of leaf infection differ among trees from true probabilities. The estimated value of 5.21 indicates some variability in the perception of infected leaves, but this variability is moderate because the shape of the underlying Beta-Binomial distribution approaches the Binomial distribution (for which perception differences are absent) (Figure 4, row 1). By contrast, the estimated value of the micro-scale fluctuation variance σ^2 (1.09) suggests a substantial variability in leaf suitability between twigs. This is evidenced by comparing the shape of the estimated Gamma-Binomial distribution with a situation with negligible differences in receptivity between twigs (Figure 4, row 2, case $\sigma^2 = 0.01$).

Table 4: Statistical summary of the inference of the parameters for the model best supported by the real data set J_{ExpP}. We used the vector of parameters θ giving the lowest AIC value in the previous model selection procedure as initial conditions of the `R` function `mle2`, to obtain maximum likelihood estimates of the vector of parameters $\hat{\theta}$ and of its matrix of variance-covariance $\hat{\sum}$. Summary statistics were derived from 1,000 random draws from the multivariate normal distribution with parameters $\hat{\theta}$ and $\hat{\sum}$ (see Appendix D.3). Columns Estimate, $q - 2.5\%$ and $q - 97.5\%$ represent the estimated value of each parameter and the quantiles 2.5% and 97.5%, respectively.

Parameter	Description	$q - 2.5\%$	Estimate	$q - 97.5\%$
r_{up}	Growth rate upstream	0.0312	0.0844	0.191
r_{dw}	Growth rate downstream	0.0114	0.0203	0.0289
λ	Mean dispersal distance	1.76	2.01	2.03
τ	Kernel exponent	0.220	0.242	0.263
γ	Tree perception	3.21	5.21	6.77
σ^2	Variance in leaf suitability	0.987	1.09	1.21

Model check consists in testing whether the selected model was indeed able –given the parameter values inferred above– to reproduce the observed data describing the epidemic wave that invaded the Durance River valley in 2008. To do so, we assessed the coverage rate of the raw sampling data (proportions of infected trees) based on their 95%-confidence intervals (Appendix D.4, Figure 5). Over all sampling dates, the mean coverage rate is high (0.74), which indicates that the model indeed captures a large part of the strong variability of the data.

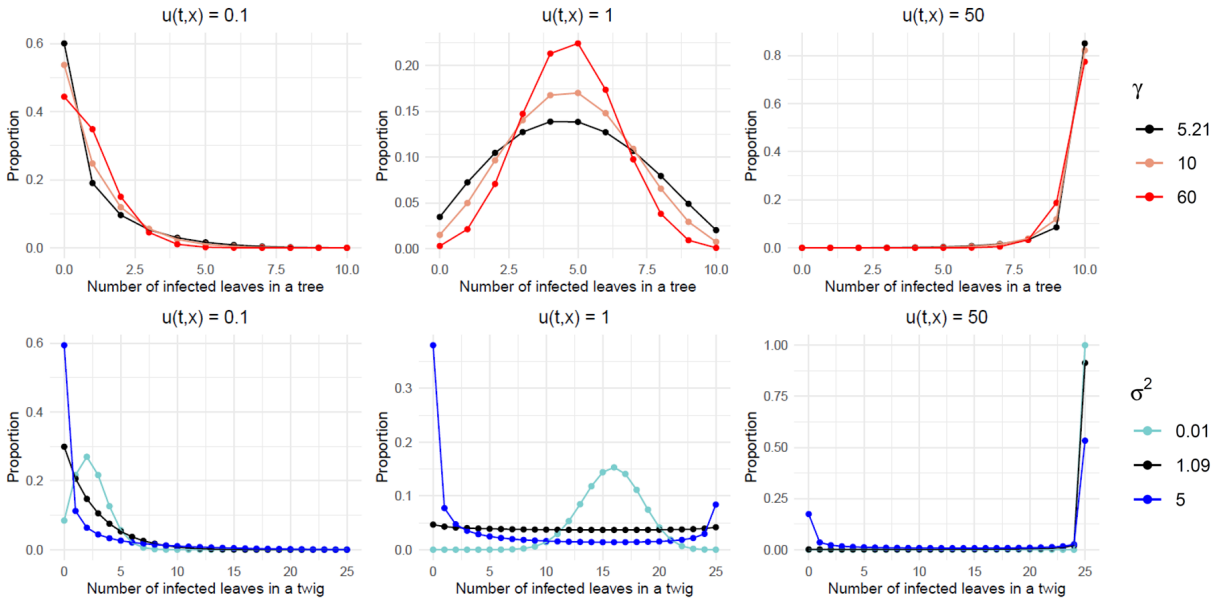

Figure 4: Distributions of the number of infected leaves in a tree and of the number of infected leaves in a twig, for increasing densities of infection $u(t,x)$, and contrasted levels of environmental heterogeneity σ^2 and γ. The number of infected leaves in a tree follows a Beta-Binomial distribution (Eq. (S8)) with $\sigma^2 = 1.09$. Its density is plotted for three tree perceptions γ: 5.21 (estimated value on the real data set), 10 (intermediate value) and 60 for which the Beta-Binomial distribution is approaching a Binomial distribution. The number of infected leaves in a twig follows a Gamma-Binomial distribution (Eq. (S14)). Its density is plotted for three leaf suitabilities σ^2: 1.09 (estimated value on the real data set), 5 (a higher value) and 0.01 a value lowering variability in leaf suitability between twigs (when σ^2 tends to 0, all twigs share the same leaf suitability).

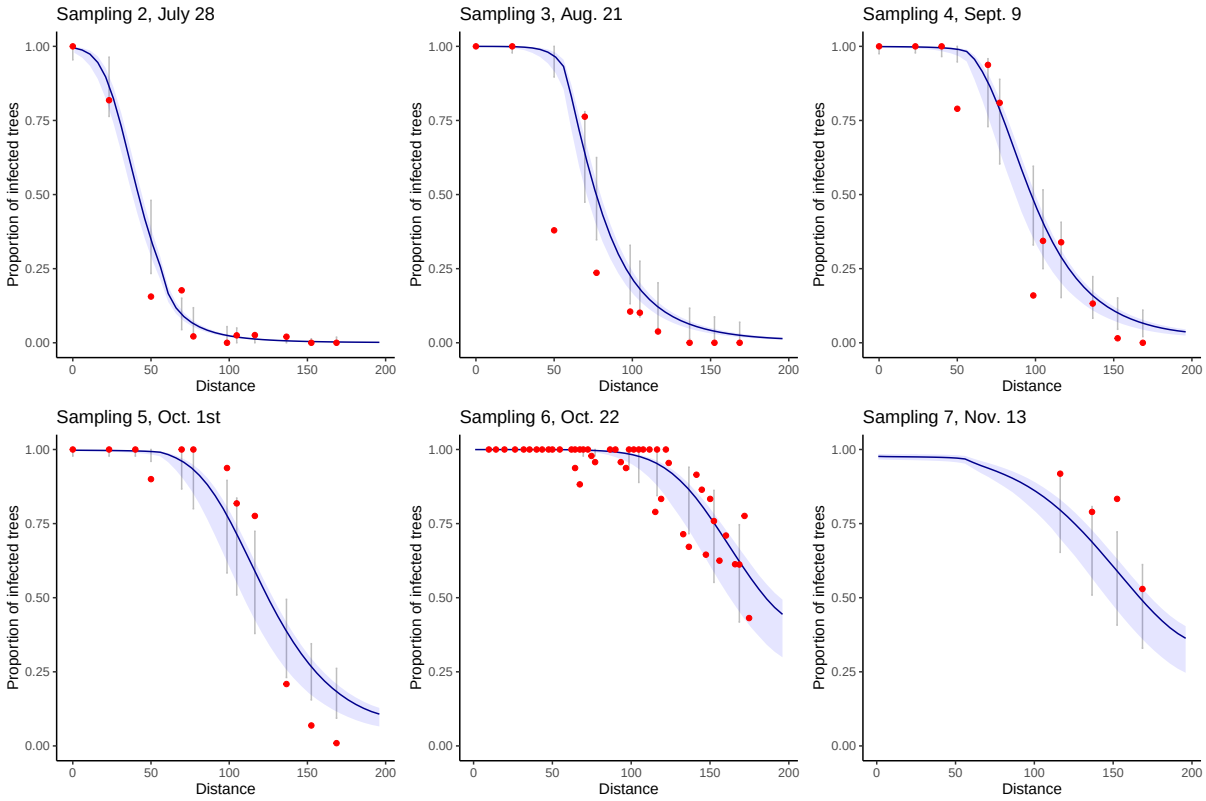

Figure 5: Model check: Coverage rates for the raw sampling. Each sampling date is represented on a separate graph. Sampling 1 is not represented because it corresponds to the initial condition of the epidemics for all simulations. Blue areas correspond to the pointwise 95% confidence envelopes for the proportion of infected trees, grey intervals correspond to the 95% prediction intervals at each site, *i.e.* taking into account the observation laws given the proportion of infected trees. Red points correspond to the observed data. Only four observations are available for sampling 7 because at this date (November 13) the leaves had already fallen from the trees located upstream the valley.

5 Discussion

This study combines mechanistic and statistical modelling to jointly infer the demographic and dispersal parameters leading to a biological invasion. A strength of the mechanistic model was to combine population growth with several dispersal processes with the aim to grasp a large diversity of processes possibly underlying invasion. The mechanistic model was coupled to a sound statistical model that considers different types of count data, which broaden the approach. Although the framework is generic, it was tuned to fit the annual spread of the poplar rust fungus *M. larici-populina* along the Durance River valley. This valley channels every year the spread of an epidemic along a one-dimensional corridor of nearly 200 km (Xhaard et al., 2012; Becheler et al., 2016). The monitoring we performed enables to build a comprehensive data set at a large spatial scale, which is mandatory to precisely infer the shape of the tail of dispersal kernels (Ferrandino, 1996; Kuparinen et al., 2007). Before applying the developed approach to our real dataset, detailed simulations were designed to prove that the demographic model can be confidently selected and its parameter values reliably inferred. A main originality in the statistical treatment of the data was to consider that habitat suitability and disease perception can vary over the sampling domain. This encapsulates a large part of the variability in the biological or observation processes and thus provides robust estimates of the parameters of biological interest.

5.1 Estimation of the dispersal kernel of the poplar rust

This study provides the first reliable estimation of the dispersal kernel of the poplar rust fungus. Dispersal kernels are firstly defined by their scale, which can be taken to correspond to the mean dispersal distance. The mean dispersal distance obtained from the best model is 2.01 km with a 95% confidence interval ranging from 1.76 to 2.27 km. Although few such estimates are available, the mean dispersal distance is likely much higher for fungal plant pathogens than for insect-transmitted pathogens. A non-systematic literature review identified only eight studies reporting dispersal kernels for plant pathogens that used data gathered in experimental designs extending over regions bigger than 1 km (Fabre et al., 2021). The mean dispersal distance of the four fungal pathosystems listed by these authors are 213 m for the ascospores of *Mycosphaerella fijiensis*, 490 m for the ascospores of *Leptosphaeria maculans*, 860 m for *Podosphaera plantaginis* and from 1380 to

381 2560 m for *Hymenoscyphus fraxineus*. Our estimates for poplar rust are in the same range as the one obtained
382 for *Hymenoscyphus fraxineus*, the causal agent of Chalara ash dieback, at regional scale (Grosdidier et al.,
383 2018).

384

385 Dispersal kernels can be further defined by their shape, which informs the fatness of their tails. We show
386 that the spread of poplar rust is best described by a fat-tailed Exponential-power kernel. The thin-tailed ker-
387 nels considered (Gaussian and exponential kernels) were clearly rejected by model selection. These results are
388 in accordance with the high dispersal ability and the long-distance dispersal events evidenced in this species
389 by population genetics analyses (Barrès et al., 2008; Becheler et al., 2016). Rust fungi are well-known to be
390 wind dispersed over long distances (Brown and Hovmøller, 2002; Aylor, 2003), some studies explaining the
391 spread of rust disease at the global scale through the air mass movement (Pan et al., 2006). Contrary to other
392 wind-borne fungi like ascomycetes (which actively project the sexually derived spores, Roper et al., 2010),
393 spores are released passively through air current, which results in both a high amount of auto-infection and fre-
394 quent events of long-distance dispersal. Recently, Severns et al. (2019) gathered experimental and simulation
395 evidence that supports that wheat stripe rust spread supports theoretical scaling relationships from power law
396 properties, another family of fat-tail dispersal kernel. Fat-tail kernels have also been identified in other plant
397 pathogens. For example, Rieux et al. (2014) and Bousset et al. (2015) found evidence for Exponential-power
398 kernels with fat-tailed for the causal agents of Black Sigatoka and Phoma stem canker, respectively. More
399 broadly, four of these eight studies listed by Fabre et al. (2021) lent support to fat-tailed kernels, including
400 plant pathogens as diverse as viruses, fungi and oomycetes. Many aerially dispersed pathogens are likely
401 to display frequent long-distance flights as soon as their life cycles include propagules. Typically, it can be
402 spore propagules but also insect vectors such as aphids escaping from plant canopy into turbulent air layer
403 (Ferrandino, 1993; Pan et al., 2010).

404

5.2 Effect of fat-tailed dispersal kernels on eco-evolutionary dynamics

The dynamics produced by the mechanistic integro-differencial models we use strongly depends on the tail of the dispersal kernel. Namely, when the equation is homogeneous (*i.e.* when the model parameters do not vary in space, leading to $r(x) = r$), it is well known that for any thin-tailed dispersal kernel J such that $\int_{\mathbb{R}} J(z)e^{\lambda|z|}dz < +\infty$ for some $\lambda > 0$, the dynamics of $u(t,x)$ is well explained using a particular solution called travelling wave. In this case, the invading front described by the solution $u(t,x)$ moves at a constant speed (Aronson and Weinberger, 1978). For a fat-tailed kernel, these particular solutions do not exist anymore and the dynamic of $u(t,x)$ describes an accelerated invasion process (Medlock and Kot, 2003; Garnier, 2011; Bouin et al., 2018). Here, we show that the dynamics of the poplar rust is better described as an accelerated invasion process rather than a front moving at a constant speed. Such accelerating wave at the epidemic front has been identified for several fungal plant pathogens dispersed by wind, including *Puccinia striiformis* and *Phytophthora infestans* the wheat stripe rust and the potato late blight, respectively (Mundt et al., 2009). However, it should be stated that fat-tailed kernels are not always associated with accelerated invasion processes. Indeed, fat-tailed kernels can be further distinguished depending on whether they are regularly varying (*e.g.* power law kernels) or rapidly varying (*e.g.* Exponential-power kernels) (Klein et al., 2006). Mathematically, it implies that power law kernels decrease even more slowly than any Exponential-power function. Biologically, fat-tailed Exponential-power kernels display rarer long-distance dispersal events than power law kernels. On the theoretical side, the kernel's properties subtly interact with demographic mechanisms such as Allee effects to possibly cancel the acceleration of invasion. With weak Allee effects (*i.e.* the growth rate is density dependent but still positive), no acceleration occurs with rapidly varying kernel whereas an acceleration could be observed for some regularly varying kernels, depending on the strength of the density dependence (Alfaro and Coville, 2017; Bouin et al., 2021). For strong Allee effects (*i.e.* a negative growth rate at low density), no acceleration can be observed for all possible kernels (Chen, 1997). On the applied side, whether or not the epidemic wave is accelerating sharply impacts the control strategies of plant pathogens (Filipe et al., 2012; Ojiambo et al., 2015; Fabre et al., 2021).

5.3 Confidence in the inference of the dispersal process

The inference framework we developed is reasonably efficient in estimating the dispersal process with frequent long-distance dispersal events as generated by Exponential-power dispersal kernels. The numerical experiments clearly show that the lower the exponent parameter τ of the Exponential-power kernel, the higher the confidence in its selection. Conversely, the identification of the dispersal process is less accurate with the Gaussian kernel. Its correct identification requires densifying the sampling. We clearly observed that integro-differential models with Gaussian dispersal kernel on the one hand and reaction-diffusion equation on the other hand are well identified with our estimation procedure when the time and space sampling is dense enough. This result may at first appear striking as a common belief tends to consider that diffusion amounts to a Gaussian dispersal kernel. However, these two models represent different movement processes (Othmer et al., 1988).

In addition, classical macroscopic diffusion, which is mainly based on Brownian motion (Othmer et al., 1988), often ignores the inherent variability among individuals' capacity of movements and as a consequence does not accurately describe the dispersal of an heterogeneous population (Hapca et al., 2009). While it is reasonable to assume that a single individual disperses via Brownian motion, this assumption hardly extents to all individuals in the population. Accordingly, we believe that integro-differential models are better suited to take into account inter-individual behaviour as the dispersal kernel explicitly models the redistribution of individuals.

5.4 Robustness and portability of the method

A strength of the approach proposed is the detailed description of the observation laws in the statistical model. The derivation of their probability density functions allows to obtain an analytical expression of the likelihood function and, as such, to use simple and fast maximum likelihood methods. However, the framework of hierarchical statistical models (Cressie et al., 2009), whose inference is often facilitated by Bayesian approaches, could likely be mobilised to improve model fit. In particular, although the coverage rate of the tree sampling was correct, it could be further improved by relaxing some hypotheses. The orange-coloured uredinia being easily seen on green leaves, we assumed that the persons in charge of the sampling

perfectly detect the disease as soon as a single uredinia is present on a leaf. However, even in this context, observation errors are likely present in our dataset as in any large spatio-temporal study. The latent variables used in hierarchical models are best suited to handle the fact that a tree observed to be healthy can actually be infected. False detection of infection could also be taken into account. This could make sense as a sister species, *M. alli-populina*, not easily discernible from *M. larici-populina* in the field, can also infect poplar leaves. This species can predominate locally in the downstream part of the Durance River valley. This could have led to over-estimate the disease severity at some locations. Yet, all infected leaves from twigs were collected and minutely inspected in the lab under a Stereo Microscope (25 magnification) to check for species identification. More generally, the statistical part of the mechanistic-statistical approaches developed could be transposed to a wide range of organisms and sampling types. The two distinct types of sampling (sampling of random leaves in trees, and of leaves grouped within twigs) apply to a wide range of species, which local distribution is aggregated into patches randomly scattered across a study site. Any biological system with two such distinct sampling types (as described in Figure 1) would fit the proposed statistical model, all the more that one can for example scale up the sampling by considering the plant (instead of the leaf) as the basic unit. Moreover, the framework naturally copes with the diversity of sampling schemes on the ground such as the absence of one sample type for all or part of the sampled sites and dates.

The mechanistic part of the model could also handle a wider diversity of hypotheses. First, the model can be adapted to take into account a wider range of dispersal kernels, such as regularly varying kernels (see above). Second, the model can also easily be adapted to take into account time and space discontinuities of its parameters. Typically, one may easily assume that the growth rate depends on daily meteorological variables. Finally, we ignore the influence of the local fluctuations of the population size on the macro-scale density of the population when stochastic fluctuations can influence epidemic dynamics Rohani et al., 2002. Here, we neglect this influence by considering that the average population size is relevant when habitat units are aggregated. Relaxing this hypothesis could be achieved by incorporating stochastic integro-differential equations. The inference of such models is currently a front of research.

5.5 Future directions

Allee effect can modulate the invasion dynamics of certain species (Dennis, 1989; Lewis and Kareiva, 1993). They could be introduced in our framework *via* a new parameter modulating the population growth depending on its size. It has been included in an earlier version of this work but led to practical identifiability problems between this Allee parameter and the population growth rates. We face a limitation caused by the type of data used. Incorporating genetic data should allow to overcome this difficulty.

As biological invasions are regularly observed retrospectively, carrying out spatio-temporal monitoring is often highly difficult, when possible. The absence of longitudinal temporal data makes model inference very difficult, in particular for its propensity to properly disentangle the effect of growth rate and dispersal. Here again, gathering genetic data may be relevant as colonisation and demographic effects generate their own specific genetic signatures (Miller et al., 2020). Furthermore, genetic data could help to identify the dispersal kernel underlying the invasion process. Indeed, the population will exhibit an erosion of its neutral diversity with a thin-tailed kernel (Edmonds et al., 2004; Hallatschek et al., 2007). Conversely, genetic diversity can be preserved all along the invasion front with a fat-tailed kernel, because of the long-distance dispersal of individuals from the back of the front, where genetic diversity is conserved (Fayard et al., 2009; Bonnefon et al., 2014).

6 Acknowledgements

We warmly thank all the collectors who participated in the monitoring of the poplar rust disease spread along the Durance River valley: Audrey Andanson, Béranger Bertin, Olivier Caël, Bénédicte Fabre, Christine Gehin, Claude Husson, Benoît Marçais, and Agathe Vialle. We also thank Benoît Marçais for fruitful discussions on disease monitoring, Bénédicte Fabre for the calculus of the density in uredinia on a poplar leaf, and Fabrice Elegbede for advices on statistical analyses. This work was supported by grants from the French National Research Agency (ANR-09-BLAN-0145, Emile project; ANR-18-CE32-0001, Clonix2D project; ANR-14-CE25-0013, project NONLOCAL, ANR-11-LABX-0002-01, Cluster of Excellence Arbre; 20-PCPA-0002, Beyond project). Constance Xhaard was supported by a PhD fellowship from the French Ministry

of Education and Research (MESR) and by Postdoc fellowship from the French National Research Agency (ANR-09-BLAN-0145, EMILE project) . Méline Saubin was supported by a PhD fellowship from INRAE and the French National Research Agency (ANR-18-CE32-0001, CLONIX2D project).

Competing interests

The authors declare that they have no known competing financial interests nor personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

Constance Xhaard, Pascal Frey, and Fabien Halkett supervised the disease monitoring. Jérôme Coville, Frédéric Fabre, Fabien Halkett, and Samuel Soubeyrand conceived and designed the study. Jérôme Coville provided a mathematical expertize on modelling long range dispersal as well as codes of simulation for the mechanistics models. Samuel Soubeyrand established the observation laws. Frédéric Fabre supervised the statistical analyses. Constance Xhaard and Fabien Halkett did preliminary analyses. Méline Saubin updated the code and did the statistical analyses. Jérôme Coville, Frédéric Fabre, Fabien Halkett, Méline Saubin, and Samuel Soubeyrand contributed to the writing of the manuscript. All authors read and approved the manuscript.

Data accessibility

`R` and `C++` scripts for model simulations and statistical analyses, as well as count data for the biological application are available on a public GitLab repository: https://gitlab.com/saubin.meline/mechanistic-statistical-model.

References

Alfaro, M. and Coville, J. (2017). Propagation phenomena in monostable integro-differential equations: Acceleration or not? *Journal of Differential Equations*, 263(9):5727–5758.

Allaire, G. (2005). *Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique*. Editions Ecole Polytechnique.

Aronson, D. G. and Weinberger, H. F. (1978). Multidimensional nonlinear diffusion arising in population genetics. *Advances in Mathematics*, 30(1):33–76.

Aylor, D. E. (2003). Spread of plant disease on a continental scale: Role of aerial dispersal of pathogens. *Ecology*, 84(8):1989–1997.

Barrès, B., Halkett, F., Dutech, C., Andrieux, A., Pinon, J., and Frey, P. (2008). Genetic structure of the poplar rust fungus *Melampsora larici-populina*: Evidence for isolation by distance in Europe and recent founder effects overseas. *Infection, Genetics and Evolution*, 8(5):577–587.

Becheler, R., Xhaard, C., Klein, E., Hayden, K. J., Frey, P., De Mita, S., and Halkett, F. (2016). Genetic signatures of a range expansion in natura: when clones play leapfrog. *Ecology and Evolution*, 6(18):6625–6632.

Berliner, L. M. (2003). Physical-statistical modeling in geophysics. *Journal of Geophysical Research*, 108(24):8776.

Bialozyt, R., Ziegenhagen, B., and Petit, R. J. (2006). Contrasting effects of long distance seed dispersal on genetic diversity during range expansion. *Journal of Evolutionary Biology*, 19(1):12–20.

Bonnefon, O., Coville, J., Garnier, J., Hamel, F., and Roques, L. (2014). The spatio-temporal dynamics of neutral genetic diversity. *Ecological Complexity*, 20:282–292.

Bouin, E., Coville, J., and Legendre, G. (2021). Sharp exponent of acceleration in general nonlocal equations with a weak Allee effect. *arXiv*, pages 1–45.

548 Bouin, E., Garnier, J., Henderson, C., and Patout, F. (2018). Thin front limit of an integro-differential
549 Fisher-KPP equation with fat-tailed kernels. *SIAM Journal on Mathematical Analysis*, 50(3):3365–3394.

550 Bourgeois, A., Gaba, S., Munier-Jolain, N., Borgy, B., Monestiez, P., and Soubeyrand, S. (2012). Inferring
551 weed spatial distribution from multi-type data. *Ecological Modelling*, 226:92–98.

552 Bousset, L., Jumel, S., Garreta, V., Picault, H., and Soubeyrand, S. (2015). Transmission of *Leptosphaeria*
553 *maculans* from a cropping season to the following one. *Annals of applied biology*, 166(3):530–543.

554 Broquet, T. and Petit, E. J. (2009). Molecular estimation of dispersal for ecology and population genetics.
555 *Annual Review of Ecology, Evolution, and Systematics*, 40(1):193–216.

556 Brown, J. K. M. and Hovmøller, M. S. (2002). Aerial dispersal of pathogens on the global and continental
557 scales and its impact on plant disease. *Science's compass*, 297:537–541.

558 Chagneau, P., Mortier, F., Picard, N., and Bacro, J. (2011). A hierarchical Bayesian model for spatial
559 prediction of multivariate non-Gaussian random fields. *Biometrics*, 67(1):97–105.

560 Chen, X. (1997). Existence, uniqueness, and asymptotic stability of traveling waves in nonlocal evolution
561 equations. *Advances in Differential Equations*, 2(1):125–160.

562 Clark, J. S., Lewis, M., and Horvath, L. (2001). Invasion by extremes: Population spread with variation in
563 dispersal and reproduction. *American Naturalist*, 157(5):537–554.

564 Clobert, J., Ims, R. A., and Rousset, F. (2004). Causes, mechanisms and consequences of dispersal. In
565 *Ecology, genetics and evolution of metapopulations*, pages 307–335. Elsevier.

566 Cressie, N., Calder, C. A., Clark, J. S., Ver Hoef, J. M., and Wikle, C. K. (2009). Accounting for uncer-
567 tainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. *Ecological*
568 *Applications*, 19(3):553–570.

569 Dennis, B. (1989). Allee effects: Population growth, critical density, and the chance of extinction. *Natural*
570 *Resource Modeling*, 3(4):481–538.

571 Dybiec, B., Kleczkowski, A., and Gilligan, C. A. (2009). Modelling control of epidemics spreading by long-
572 range interactions. *Journal of the Royal Society Interface*, 6(39):941–950.

573 Edmonds, C. A., Lillie, A. S., and Cavalli-Sforza, L. L. (2004). Mutations arising in the wave front of an
574 expanding population. *Proceedings of the National Academy of Sciences*, 101(4):975–979.

575 Fabre, F., Coville, J., and Cunniffe, N. J. (2021). Optimising reactive disease management using spatially
576 explicit models at the landscape scale. In Scott, P. R., Strange, R. N., Korsten, L., and Gullino, M. L.,
577 editors, *Plant disease and food security in the 21st century*, pages 47–72. Springer International Publishing.

578 Fayard, J., Klein, E., and Lefèvre, F. (2009). Long distance dispersal and the fate of a gene from the
579 colonization front. *Journal of Evolutionary Biology*, 22(11):2171–2182.

580 Ferrandino, F. J. (1993). Dispersive epidemic waves: I. Focus expansion within a linear planting. *Phytopath-*
581 *ology*, 83(8):795.

582 Ferrandino, F. J. (1996). Lenght scale of disease spread: Fact or artifact of experimental geometry. *Phyto-*
583 *pathology*, 86:806–811.

584 Fife, P. C. (1996). An integrodifferential analog of semilinear parabolic PDEs. In *Partial differential equations*
585 *and applications*, volume 177 of *Lecture Notes in Pure and Appl. Math.*, pages 137–145. Dekker, New York.

586 Filipe, J. A. N., Cobb, R. C., Meentemeyer, R. K., Lee, C. A., Valachovic, Y. S., Cook, A. R., Rizzo, D. M.,
587 and Gilligan, C. A. (2012). Landscape epidemiology and control of pathogens with cryptic and long-distance
588 dispersal: Sudden Oak death in northern Californian forests. *PLoS Computational Biology*, 8(1):e1002328.

589 Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of Eugenics*, 7:355–369.

590 Gandon, S. and Michalakis, Y. (2002). Local adaptation, evolutionary potential and host parasite coevolution:
591 Interactions between migration, mutation, population size and generation time. 15(1):451–462.

592 Garnier, J. (2011). Accelerating solutions in integro-differential equations. *SIAM J. Math. Anal.*, 43:1955–
593 1974.

Georgescu, V., Desassis, N., Soubeyrand, S., Kretzschmar, A., and Senoussi, R. (2014). An automated MCEM algorithm for hierarchical models with multivariate and multitype response variables. *Communications in Statistics - Theory and Methods*, 43(17):3698–3719.

Gotway, C. A. and Young, L. J. (2002). Combining incompatible spatial data. *Journal of the American Statistical Association*, 97(458):632–648.

Grosdidier, M., Ioos, R., Husson, C., Cael, O., Scordia, T., and Marçais, B. (2018). Tracking the invasion: dispersal of *Hymenoscyphus fraxineus* airborne inoculum at different scales. *FEMS microbiology ecology*, 94(5):1–11.

Hallatschek, O. and Fisher, D. S. (2014). Acceleration of evolutionary spread by long-range dispersal. *Proceedings of the National Academy of Sciences*, 111(46):E4911–E4919.

Hallatschek, O., Hersen, P., Ramanathan, S., and Nelson, D. R. (2007). Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences*, 104(50):19926–19930.

Hapca, S., Crawford, J. W., and Young, I. M. (2009). Anomalous diffusion of heterogeneous populations characterized by normal diffusion at the individual level. *Journal of the Royal Society Interface*, 6(30):111–122.

Hefley, T. J., Hooten, M. B., Russell, R. E., Walsh, D. P., and Powell, J. A. (2017). When mechanism matters: Bayesian forecasting using models of ecological diffusion. *Ecology Letters*, 20(5):640–650.

Hutson, V., Martinez, S., Mischaikow, K., and Vickers, G. T. (2003). The evolution of dispersal. *Journal of Mathematical Biology*, 47(6):483–517.

Ibrahim, K. M., Nichols, R. A., and Hewitt, G. M. (1996). Spatial patterns of genetic variation generated by different forms of dispersal during range expansion. *Heredity*, 77:282–291.

Klein, E., Lavigne, C., Picault, H., Renard, M., and Gouyon, P. H. (2006). Pollen dispersal of oilseed rape: Estimation of the dispersal function and effects of field dimension. *Journal of Applied Ecology*, 43(1):141–151.

Kolmogorov, A. N., Petrovsky, I. G., and Piskunov, N. S. (1937). Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Bulletin Université d'État à Moscow (Bjul. Moskowskogo Gos. Univ)*, pages 1–26.

Kot, M., Lewis, M. A., and van den Driessche, P. (1996). Dispersal data and the spread of invading organisms. *Ecology*, 77(7):2027–2042.

Kuparinen, A., Snäll, T., Vänskä, S., and O'Hara, R. B. (2007). The role of model selection in describing stochastic ecological processes. *Oikos*, 116(6):966–974.

Lewis, M. A. and Kareiva, P. (1993). Allee dynamics and the spread of invading organisms. *Theoretical Population Biology*, 42:141–158.

Louvrier, J., Papaïx, J., Duchamp, C., and Gimenez, O. (2020). A mechanisticstatistical species distribution model to explain and forecast wolf (*Canis lupus*) colonization in South-Eastern France. *Spatial Statistics*, 36:100428.

Macdonald, D. W. and Johnson, D. D. P. (2001). Dispersal in theory and practice: consequences for conservation biology. In Clober, T. J., Danchin, E., Dhondt, A. A., and Nichols, J. D., editors, *Dispersal*, chapter 25, pages 361–374. Oxford University Press, Oxford, UK.

Maupetit, A., Larbat, R., Pernaci, M., Andrieux, A., Guinet, C., Boutigny, A. L., Fabre, B., Frey, P., and Halkett, F. (2018). Defense compounds rather than nutrient availability shape aggressiveness trait variation along a leaf maturity gradient in a biotrophic plant pathogen. *Frontiers in Plant Science*, 9(September).

Medlock, J. and Kot, M. (2003). Spreading disease: Integro-differential equations old and new. *Math. Biosci.*, 184(2):201–222.

Miller, T. E. X., Angert, A. L., Brown, C. D., Lee-Yaw, J. A., Lewis, M., Lutscher, F., Marculis, N. G., Melbourne, B. A., Shaw, A. K., Szcs, M., Tabares, O., Usui, T., Weiss-Lehman, C., and Williams, J. L. (2020). Eco-evolutionary dynamics of range expansion. *Ecology*, 101(10):1–14.

Mollison, D. (1977). Spatial contact models for ecological and epidemic spread. *J. R. Stat. Ser. B Stat. Methodol.*, 39:283–326.

643 Mundt, C. C., Sackett, K. E., Wallace, L. D., Cowger, C., and Dudley, J. P. (2009). Long-distance dispersal
644 and accelerating waves of disease: Empirical relationships. *American Naturalist*, 173(4):456–466.

645 Murray, J. D. (2002). *Mathematical Biology*, volume 17. Springer-Verlag, third edition.

646 Nathan, R. (2001). The challenges of studying dispersal. *Trends in Ecology and Evolution*, 16(9):481–483.

647 Nathan, R., Klein, E., Robledo-Arnuncio, J. J., and Revilla, E. (2012). 15 - Dispersal kernels: Review. In
648 Clobert, J., Baguette, M., Benton, T. G., and Bullock, J. M., editors, *Dispersal ecology and evolution*, pages
649 186–210. Oxford.

650 Nathan, R., Perry, G., Cronin, J. T., Strand, A. E., and Cain, M. L. (2003). Methods for estimating long-
651 distance dispersal. *Nature ecology & evolution*, 103(2):261–273.

652 Nembot Fomba, C. G., Ten Hoopen, G. M., Soubeyrand, S., Roques, L., Ambang, Z., and Takam Soh, P.
653 (2021). Parameter estimation in a PDE model for the spatial spread of cocoa black pod disease. *Bulletin*
654 *of Mathematical Biology*, 83:1–28.

655 Nichols, R. A. and Hewitt, G. M. (1994). The genetic consequences of long distance dispersal during coloniz-
656 ation. *Heredity*, 72:312–317.

657 Ojiambo, P. S., Gent, D. H., Quesada-Ocampo, L. M., Hausbeck, M. K., and Holmes, G. J. (2015). Epidemi-
658 ology and population biology of Pseudoperonospora cubensis : A model system for management of downy
659 mildews. *Annual Review of Phytopathology*, 53(1):223–246.

660 Okubo, A. and Levin, S. A. (2002). *Diffusion and Ecological Problems – Modern Perspectives*. Second edition,
661 Springer-Verlag, New York.

662 Othmer, H. G., Dunbar, S. R., and Alt, W. (1988). Models of dispersal in biological systems. *Journal of*
663 *mathematical biology*, 26(3):263–298.

664 Pan, Z., Li, X., Yang, X. B., Andrade, D., Xue, L., and McKinney, N. (2010). Prediction of plant diseases
665 through modelling and monitoring airborne pathogen dispersal. *CAB Reviews: Perspectives in Agriculture,*
666 *Veterinary Science, Nutrition and Natural Resources*, 5(018).

667 Pan, Z., Yang, X. B., Pivonia, S., Xue, L., Pasken, R., and Roads, J. (2006). Long-term prediction of soybean
668 rust entry into the continental United States. *Plant Disease*, 90(7):840–846.

669 Papaïx, J., Soubeyrand, S., Bonnefon, O., Walker, E., Louvrier, J., Klein, E., and Roques, L. (2022). Inferring
670 mechanistic models in spatial ecology using a mechanisticstatistical approach. In *Statistical Approaches for*
671 *Hidden Variables in Ecology*, pages 69–95. Wiley.

672 Petit, R. J. (2004). Biological invasions at the gene level. *Diversity and Distributions*, 10(3):159–165.

673 Petit, R. J. (2011). Early insights into the genetic consequences of range expansions. *Heredity*, 106:203–204.

674 R Core Team (2018). R: A language and environment for statistical computing.

675 Rieux, A., Lenormand, T., Carlier, J., De Lapeyre de Bellaire, L., and Ravigné, V. (2013). Using neutral cline
676 decay to estimate contemporary dispersal: A generic tool and its application to a major crop pathogen.
677 *Ecology Letters*, 16(6):721–730.

678 Rieux, A., Soubeyrand, S., Bonnot, F., Klein, E. K., Ngando, J. E., Mehl, A., Ravigne, V., Carlier, J.,
679 and Bellaire, L. (2014). Long-distance wind-dispersal of spores in a fungal plant pathogen: Estimation of
680 anisotropic dispersal kernels from an extensive field experiment. *PLoS ONE*, 9(8):e103225.

681 Ritz, C., Baty, F., Streibig, J. C., and Gerhard, D. (2015). Dose-Response analysis using R. *PLOS ONE*,
682 10(12).

683 Rohani, P., Keeling, M. J., and Grenfell, B. T. (2002). The interplay between determinism and stochasticity
684 in childhood diseases. *The American Naturalist*, 159:469–481.

685 Roper, M., Seminara, A., Bandi, M. M., Cobb, A., Dillard, H. R., and Pringle, A. (2010). Dispersal of fungal
686 spores on a cooperatively generated wind. *Proceedings of the National Academy of Sciences of the United*
687 *States of America*, 107(41):17474–17479.

688 Roques, L., Soubeyrand, S., and Rousselet, J. (2011). A statistical-reaction-diffusion approach for analyzing
689 expansion processes. *Journal of Theoretical Biology*, 274(1):43–51.

690 Roques, L., Walker, E., Franck, P., Soubeyrand, S., and Klein, E. K. (2016). Using genetic data to estimate
691 diffusion rates in heterogeneous landscapes. *Journal of Mathematical Biology*, 73:397–422.

692 Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., Mcroberts, N., and Nelson, A. (2019). The global
693 burden of pathogens and pests on major food crops. *Nature ecology & evolution*, 3(3):430–439.

694 Severns, P. M., Sackett, K. E., Farber, D. H., and Mundt, C. C. (2019). Consequences of long-distance
695 dispersal for epidemic spread: Patterns, scaling, and mitigation. *Plant Disease*, 103(2):177–191.

696 Shigesada, N. and Kawasaki, K. (1997). *Biological invasions: Theory and practice*. Oxford University Press,
697 UK.

698 Soubeyrand, S., de Jerphanion, P., Martin, O., Saussac, M., Manceau, C., Hendrikx, P., and Lannou, C.
699 (2018). Inferring pathogen dynamics from temporal count data: the emergence of *Xylella fastidiosa* in
700 France is probably not recent. *New Phytologist*, 219:824–836.

701 Soubeyrand, S., Enjalbert, J., Sanchez, A., and Sache, I. (2007). Anisotropy, in density and in distance,
702 of the dispersal of yellow rust of wheat: experiments in large field plots and estimation. *Phytopathology*,
703 97:1315–1324.

704 Soubeyrand, S., Laine, A. L., Hanski, I., and Penttinen, A. (2009a). Spatio-temporal structure of host-
705 pathogen interactions in a metapopulation. *The American Naturalist*, 174(3):308–320.

706 Soubeyrand, S., Neuvonen, S., and Penttinen, A. (2009b). Mechanical-statistical modeling in ecology: From
707 outbreak detections to pest dynamics. *Bull. Math. Bio.*, 71:318–338.

708 Soubeyrand, S. and Roques, L. (2014). Parameter estimation for reaction-diffusion models of biological
709 invasions. *Population Ecology*, 56(2):427–434.

710 Soubeyrand, S., Sache, I., Hamelin, F., and Klein, E. K. (2015). Evolution of dispersal in asexual populations:
711 to be independent, clumped or grouped? *Evolutionary Ecology*, 29:947–963.

712 Strange, R. N. and Scott, P. R. (2005). Plant disease: A threat to global food security. *Annual Review of*
713 *Phytopathology*, 43:83–116.

714 Szymańska, Z., Skrzeczkowski, J., Miasojedow, B., and Gwiazda, P. (2021). Bayesian inference of a non-local
715 proliferation model. *Royal Society Open Science*, 8(11).

716 Wikle, C. K. (2003a). Hierarchical Bayesian models for predicting the spread of ecological processes. *Ecology*,
717 84(6):1382–1394.

718 Wikle, C. K. (2003b). Hierarchical models in environmental science. *International Statistical Review*,
719 71(2):181–199.

720 Xhaard, C., Barrès, B., Andrieux, A., Bousset, L., Halkett, F., and Frey, P. (2012). Disentangling the
721 genetic origins of a plant pathogen during disease spread using an original molecular epidemiology approach.
722 *Molecular Ecology*, 21(10):2383–2398.