A mechanistic-statistical approach to infer dispersal and demography from invasion dynamics, applied to a plant pathogen

Méline Saubin~~[1]~~[1*], Jérome Coville[2], Constance Xhaard[1,2,3], Pascal Frey[1], Samuel Soubeyrand[2], Fabien Halkett~~[1]~~ [1#] and Frédéric Fabre~~[4]~~[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

~~**Corresponding author**~~[*] **Current adress:** ~~Meline Saubin~~ ~~**Current adress:**~~ Populationsgenetik, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising, Germany

~~**E-mail:**~~[#] These authors co-directed this work.

Corresponding authors:

Méline Saubin (meline.saubin@tum.de)

Fabien Halkett (fabien.halkett@inrae.fr)

Frédéric Fabre (frederic.fabre@inrae.fr)

Keywords: ~~1-D colonisation, dispersal kernel, long-distance dispersal, multiple data types, population dynamic, spatio-temporal model~~

Abstract

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 for 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. Here, we propose a mechanistic-statistical framework to estimate dispersal from 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 latter 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, clumped or not. As such, we take into account the variability in both habitat suitability and occupancy perception. 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~~ correlation coefficient > 0.9 between true and fitted values)~~, except for occupancy perception (Correlation Coefficient $= 0.83 - 0.85$)~~. The dispersal process that is the most confidently identified is Exponential-Power (*i.e.* fat-tailed) kernel. 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~~ 1.94 km and an exponent parameter of 0.24 characterising a fat-tailed kernel with frequent LDD events. By considering the whole range of possible dispersal processes our method forms a robust inference framework. It can be employed for a variety of organisms, provided they are monitored in time and space along a one-dimension invasion.

Keywords: 1-D colonisation, dispersal kernel, long-distance dispersal, multiple data types, population dynamic, spatio-temporal model

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). The 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 (Klein et al., 2006). Short-tailed kernels (also referred to as thin-tailed) generate an invasion front of constant velocity, whereas long-tailed kernels (also referred to as fat-tailed) can cause an accelerating front of colonisation (Ferrandino, 1993; Kot et al., 1996; Clark et al., 2001; Mundt et al., 2009; Hallatschek and Fisher, 2014). Long-tailed kernels, characterised by more frequent long-distance dispersal (LDD) events than an exponential kernel that shares the same mean dispersal distance, 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). In that quest, mechanistic-statistical models 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) while allowing 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. 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 following trajectories modelled using a Brownian motion or a more general stochastic diffusion process. Despite their long standing history, the incorporation of reaction-diffusion equations into mechanistic-statistical approaches to estimate parameters of interest from spatio-temporal data essentially dates back to the early 2000s (*e.g.* Wikle, 2003a; Soubeyrand and Roques, 2014; Louvrier et al., 2020; Nembot Fomba et al., 2021). By contrast to reaction-diffusion equations, integro-differential equations encode trajectories modelled by jump diffusion processes and 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 organism's dispersal process. In the presence of long-distance dispersal, the biological interpretation of the estimated diffusion parameters with an R.D. equation would be misleading. However, integro-differential equations are numerically more demanding to simulate than reaction-diffusion equations. 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. 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). This way of sampling is all the more interesting as the probabilistic model describing the observation process in the mechanistic-statistical approach can handle such multiple datasets (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 a 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 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. 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 the sampling 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 fungal 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

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.):

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

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, $r(x)u(t,x)\left(1 - \dfrac{u(t,x)}{K}\right)$ 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.). To cover a large spectrum of possible dispersal processes, we use the following parametric form for the kernel J:

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

with mean dispersal distance $\lambda := \alpha\dfrac{\Gamma\left(\frac{2}{\tau}\right)}{\Gamma\left(\frac{1}{\tau}\right)}$. Varying the value of τ leads to the 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 ($\tau < 1, \lambda = \alpha\Gamma\left(\dfrac{2}{\tau}\right)/\Gamma\left(\dfrac{1}{\tau}\right)$). 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 S1.

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). It is worth considering both scales 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 locally (*e.g.* because of variations in the micro-climate and the host susceptibility). Such local fluctuations are accounted for by a conditional probability distribution on $u(t,x)$, the macro-scale population density, which depends on the (unobserved) suitability of the habitat unit as follow. 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 every 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:

$$N_i(t) \mid u(t,x_i), R_i(t) \sim \text{Poisson}(u(t,x_i)R_i(t)), \tag{2}$$

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:

$$R_i(t) \sim \text{Gamma}(\sigma^{-2}, \sigma^2). \tag{3}$$

This parametrisation 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:

$$N_i(t) \mid u(t,x_i) \sim \text{Negative-Binomial}\left(\sigma^{-2}, \underline{\frac{u(t,x_i)\sigma^2}{1+u(t,x_i)\sigma^2}}\,\frac{1}{1+u(t,x_i)\sigma^2}\right). \tag{4}$$

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 (Figure 1). We consider that the habitat unit is a leaf. The fungal population is monitored in sampling sites $s \in \{1,\dots,S\}$ and at sampling times $t \in \{t_1,\dots,t_K\}$. 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. 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. Each observed tree and twig are assumed to be observed only once during the sampling period. Therefore, habitat suitabilities $R_i(t)$ are considered independent in time.

The raw sampling is focused on trees, considered as a group of independent leaves regarding their suitabilities. This assumption can be made if the leaves observed on the same tree are sufficiently far from each other and represent a large variety of environmental conditions, and therefore habitat suitabilities (for example, leaves observed all around a tree will not have the same sun exposition, nor the same humidity depending on their height and their relative positions to the trunk). In each sampling site s and at each sampling time t, a number B_{st} of trees are 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 Binomial distribution f_{st}^{raw} described in Appendix S2.2. Its success probability depends on the variabilities of (i) the biological process through the variance parameter σ^2 of habitat suitabilities, and (ii) the observation process through a parameter γ. This parameter describes 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 (2–3). 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 S2.3. The distribution f_{st}^{ref} is a new mixture distribution (called Gamma-Binomial distribution) obtained using Equations (2–3) and taking into account the spatial dependence and the variance parameter of unobserved suitabilities (see Appendix S2.3).

This sampling scheme and its vocabulary (leaves, twigs and trees) are 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 fits a two-stage sampling as presented in Figure 1.

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:

$$L(\theta) = \prod_{s=1}^{S} \prod_{t=t_1}^{t_K} \left\{ f_{st}^{\text{raw}}(Y_{st}) \left(\prod_{g=1}^{G_{st}} f_{st}^{\text{ref}}(Y_{stg}) \right)^{\mathbb{1}(Y_{st} > \bar{y})} \right\}, \tag{5}$$

where $\mathbb{1}(\cdot)$ denotes the indicator function and expressions of f_{st}^{raw} and f_{st}^{ref} adapted to the case study tackled below are given by Equations (S14) and (S18) in Appendix S2. The power $\mathbb{1}(Y_{st} > \bar{y})$ equals to 1 if $Y_{st} > \bar{y}$ and 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 S2.2 can be made to obtain Equation (5), 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 S4), 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 ~~$n = 100$~~ $n = 160$ times. Details on the simulation procedure, the conditions used to generate realistic invasions, and on the estimation algorithm are provided in Appendix S4.1. Practical identifiability was tested by means of correlation coefficients between the true and estimated parameter values (see Table 1, Appendix S2: 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.~~0.98 (except for mean dispersal distance λ under R.D., correlation coefficient of 0.94). 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~~0.93.

The parameter defining the micro-scale fluctuations, σ^2, leads 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~~0.91 to 0.97).

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 (Appendix S4.2). The model selection procedure is the most efficient for the dispersal hypotheses Exponential-power J_{ExpP}, ~~Exponential J_{Exp}, and reaction-diffusion R.D., with 70%, 62% and 58%~~ with 78% 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%~~ 17% 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~~ 217.50 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~~ 0.76 point on average). ~~Model selection does not allow to correctly select the~~ The model selection is the least efficient for the Gaussian kernel J_{Gauss} (Table 2)~~. Indeed~~, with only ~~26%~~ 45% 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~~ improved to 80% by densifying the sampling scheme (Appendix S4.5: Table S2). Finally, note that when the invasion process is simulated under model R.D. or J_{Gauss}, a short-tail kernel is almost always selected and, thus, never confounded with the fat-tail kernel J_{ExpP}.

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 that 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 fungal infection at time t at point x on a poplar leaf. Each leaf has a carrying capacity of 750 fungal infections (Appendix S5).

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 larches coexist (Prelles), and progresses along the valley (Becheler et al., 2016). In autumn, 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 S4 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 to 2m high) growing on the stands by the riverside.

Two monitorings were conducted, corresponding to the raw and refined sampling, as described in previous sections. For the raw sampling, we prospected each site at each date to search for rust disease by inspecting randomly distributed poplar trees (different trees at different dates for a given site). Depending on rust incidence and poplar tree accessibility, 40 to 150 trees (mean 74) were checked for disease. Each tree was inspected through a global scan of the leaves on different twigs until at least one infected leaf was found or after 30 s of inspection. The tree was denoted infected or healthy, respectively. This survey method amounts 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 S3). 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

Model selection was used to decipher which dispersal process was best supported by the data set for five initial parameter values. 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 parameter values lead to similar estimations. ~~Under the R.D. hypothesis, however, initial parameter values 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.~~

The estimation of the parameters for the best model along with their confidence intervals (Appendix S4.3) are summarised in Table 4. The parameters of the Exponential-power kernel firstly indicate that the mean distance travelled by rust spores is estimated at ~~2.01~~ 1.94 km. Second, its mean exponent parameter τ 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~~0.085 and 0.023, respectively. The estimate of the parameter of the observation model, γ, is ~~5.21~~4.82. This parameter represents how perceived probabilities of leaf infection differ among trees from true probabilities. The estimated value of ~~5.21~~ 4.82 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~~1.29) 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 $\sigma^2 = 0.01$).

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 S4.4, Figure 5). Over all sampling dates, the total coverage rate is high (~~0.75~~0.71), which indicates that the model indeed captures a large part of the strong variability of the data. By comparison, coverage rates given by models J_{Exp} and J_{Gauss} (~~0.69~~ 0.68 and 0.67, respectively) show a poorer fit to the data, especially for the first sampling date (Figures S6, S7) where the epidemic intensity is underestimated upstream and overestimated downstream.

5 Discussion

This study combines mechanistic and statistical modelling to jointly infer the demographic and dispersal parameters underlying a biological invasion. A strength of the mechanistic model was to combine population growth with a large diversity of dispersal processes. The mechanistic model was coupled to a sound statistical model that considers different types of count data. These observation laws consider that habitat suitability and disease perception can vary over the sampling domain. Simulations were designed to prove that the demographic model can be confidently selected and its parameter values reliably inferred. 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). A widely used alternative to the mechanistic-statistical approaches is to consider purely correlative approaches. However, the estimated parameters defining the strength of the temporal and spatial dependencies (as estimated for example using R-INLA package approach, Rue et al., 2009) will not allow to distinguish between the different shapes of dispersal kernels, which was the main goal of our work.

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~~ 1.94 km with a 95% confidence interval ranging from ~~1.76 to 2.27~~ 1.78 to 2.12 km. 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 distances of the four fungal pathosystems listed by these authors are 213 m for the ascospores of *Mycosphaerella fijiensis* (Rieux et al., 2014), 490 m for the ascospores of *Leptosphaeria maculans* (Bousset et al., 2015), 860 m for *Podosphaera plantaginis* (Soubeyrand et al., 2009a) and from 1380 to 2560 m for *Hymenoscyphus fraxineus* (Grosdidier et al., 2018). Our estimates for poplar rust are in the same range as the one obtained at regional scale for *Hymenoscyphus* *fraxineus*, the causal agent of Chalara ash dieback (Grosdidier et al., 2018).

Dispersal kernels can be further defined by their shape. We show that the spread of poplar rust is best described by a fat-tailed Exponential-power kernel. The thin-tailed kernels considered (Gaussian and exponential kernels) were clearly rejected by model selection. These results are in accordance with the high dispersal ability and the long-distance dispersal events evidenced in this species by population genetics analyses (Barrès et al., 2008; Becheler et al., 2016). Rust fungi are well-known to be wind dispersed over long distances (Brown and Hovmøller, 2002; Aylor, 2003). Recently, Severns et al. (2019) gathered experimental and simulation evidence that supports that wheat stripe rust spread supports theoretical scaling relationships from power law properties, another family of fat-tail dispersal kernel. In fact, many aerially dispersed pathogens are likely to display frequent long-distance flights as soon as their propagules (spores, insect vectors) escape from plant canopy into turbulent air layer (Ferrandino, 1993; Pan et al., 2010). Accordingly, four of these eight studies listed by Fabre et al. (2021) lent support to fat-tailed kernels, including plant pathogens as diverse as viruses, fungi, and oomycetes.

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). By contrast, 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 kernels 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 thin-tail kernels. The requirement for improving the capacity to distinguish between thin-tail kernels may lie in the sampling scheme. Here, our sampling sites are regularly spaced, over a large sampling domain of 200 km, which is better suited to monitor long-distance dispersal (Kuparinen et al., 2007). Sampling schemes with more frequent data in both time and space (or nested spatial sampling) might improve kernel identification.

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 at the population scale (Hapca et al., 2009). While it is reasonable to assume that a single individual disperses via Brownian motion, this assumption hardly extends 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. Model inference was however not straightforward due to local optimum issues. In order to achieve satisfying computational efficiency, we developed an *ad hoc* hybrid strategy initiated from 20 initial values and combining the two classical Nelder-Mead and Nlminb optimisation algorithms. 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. Sharing the sampling effort between raw and refined samples improves the estimations. 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 study with two such distinct sampling types (as described in Figure 1) would fit the proposed statistical model. 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. Finally, we used the first sampling date to estimate independently the initial population densities $u(0,x)$ that were then fixed among all simulated epidemics. Future works could as well jointly estimate $u(0,x)$ as part of θ.

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 parameter heterogeneity in time and space of its parameters. Similarly, 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

As biological invasions are regularly observed retrospectively, carrying out spatio-temporal monitoring is often highly difficult, when possible. A small number of longitudinal temporal data makes model inference very difficult, in particular for its propensity to properly disentangle the effect of growth rate and dispersal. Incorporating genetic data into the framework proposed here is a challenge that must be met to get around this problem. Indeed, colonisation and demographic effects such as Allee effect generate their own specific genetic signatures (Dennis, 1989; Lewis and Kareiva, 1993; Miller et al., 2020). Similarly, genetic data could help to identify the dispersal kernel underlying the invasion process, as 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).

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).

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 expertise on modelling long-range dispersal as well as codes of simulation for the mechanistic 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.

Competing interests

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

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 Zenodo repository (DOI:10.5281/zenodo.7906841), extracted from 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.

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.

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

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

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

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

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

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

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

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

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

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

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

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

Fabre, F., Coville, J., and Cunniffe, N. J. (2021). Optimising reactive disease management using spatially explicit models at the landscape scale. In Scott, P. R., Strange, R. N., Korsten, L., and

Gullino, M. L., editors, *Plant disease and food security in the 21st century*, pages 47–72. Springer International Publishing.

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

Ferrandino, F. J. (1993). Dispersive epidemic waves: I. Focus expansion within a linear planting. *Phytopathology*, 83(8):795.

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

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

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

Gandon, S. and Michalakis, Y. (2002). Local adaptation, evolutionary potential and host parasite coevolution: Interactions between migration, mutation, population size and generation time. *Journal of Evolutionary Biology*, 15(3):451–462.

Garnier, J. (2011). Accelerating solutions in integro-differential equations. *SIAM Journal on Mathematical Analysis*, 43:1955–1974.

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 mechanistic-statistical 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.

Medlock, J. and Kot, M. (2003). Spreading disease: Integro-differential equations old and new. *Mathematical Biosciences*, 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. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39:283–326.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

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

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

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

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

Soubeyrand, S., Neuvonen, S., and Penttinen, A. (2009b). Mechanical-statistical modeling in ecology: From outbreak detections to pest dynamics. *Bulletin of Mathematical Biology*, 71:318–338.

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

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

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

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

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

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

Tables

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~~ 160 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_{dw}	Growth rate downstream	~~0.99(1.10^{-3})~~ 0.997 (0.001)	~~0.99(1.10^{-3})~~ 0.997 (0.001)	~~0.99(2.1~~
ω	Growth rate modulator	~~0.99($< 10^{-3}$)~~ 0.997 (0.001)	~~0.99($< 10^{-3}$)~~ 0.992 (0.003)	~~0.99(1.1~~
λ	Mean dispersal distance	~~0.99(5.10^{-3})~~ 0.983 (0.007)	~~0.98(8.10^{-3})~~ 0.993 (0.004)	~~0.99(1.1~~
τ	Kernel exponent	NA	NA	~~0.95(1.1~~
γ	Tree perception	~~0.85(4.10^{-2})~~ 0.969 (0.006)	~~0.83(4.10^{-2})~~ 0.966 (0.006)	~~0.83(5.1~~
σ^2	Variance in leaf suitability	~~0.99(1.10^{-3})~~ 0.996 (0.001)	~~0.99($< 10^{-3}$)~~ 0.997 (0.001)	~~0.99(<~~

Table 2: Efficiency of model selection using Akaike information criterion (AIC). The four first columns indicate the proportion of cases, among ~~50~~ 60 replicates, where each tested model was selected using AIC, given that data sets were generated under a particular model (*i.e.* true model). Column $d\mathrm{AIC}_{\mathrm{true}}$ (*resp.* $d\mathrm{AIC}_{\mathrm{wrong}}$) 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.	$d\mathrm{AIC}_{\mathrm{true}}$	$d\mathrm{AIC}_{\mathrm{wrong}}$
J_{Exp}	**~~0.62~~0.50**	~~0.22~~ 0.32	~~0.06~~ 0.05	~~0.10~~ 0.13	~~0.84~~ 0.90	~~0.74~~1.10
J_{Gauss}	~~0.34~~ 0.25	**~~0.26~~0.45**	~~0.00~~ 0.02	~~0.40~~ 0.28	~~1.08~~ 1.30	~~0.55~~0.67
J_{ExpP}	~~0.20~~ 0.17	~~0.04~~ 0.05	**~~0.70~~0.78**	~~0.06~~ 0	**~~89.62~~217.50**	~~0.38~~0.76
R.D.	0.22	0.18	~~0.24~~ 0	~~0.00~~ **0.60**	**~~0.58~~** 2.37	~~0.71~~ ~~0.23~~0.33

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. $\text{AIC}_{\text{median}}$ and AIC_{sd} represent the median and standard deviation among the AIC obtained from five initial parameter values.

Dispersal	$\text{AIC}_{\text{median}}$	AIC_{sd}
J_{Exp}	~~5476~~ 5461	~~0.68~~ 5.81
J_{Gauss}	~~5510~~ 5493	~~1.03~~ 0.15
J_{ExpP}	**~~5179~~5163**	**~~1.32~~0.01**
R.D.	~~6303~~ 6190	~~655.60~~ 0.03

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 parameter values 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 S4.3). 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~~ 0.0786	~~0.0844~~ 0.0853	~~0.191~~ 0.0897
r_{dw}	Growth rate downstream	~~0.0114~~ 0.0143	~~0.0203~~ 0.0230	~~0.0289~~ 0.0311
λ	Mean dispersal distance	~~1.76~~ 1.78	~~2.01~~ 1.94	~~2.03~~ 2.12
τ	Kernel exponent	~~0.220~~ 0.226	~~0.242~~ 0.243	~~0.263~~ 0.260
γ	Tree perception	~~3.21~~ 3.49	~~5.21~~ 4.82	~~6.77~~ 6.11
σ^2	Variance in leaf suitability	~~0.987~~ 1.23	~~1.09~~ 1.29	~~1.21~~ 1.36

Figures

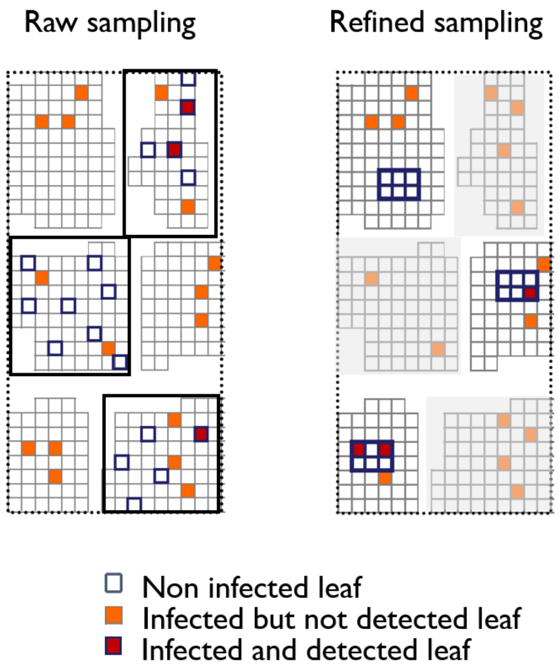
Raw sampling
Refined sampling
□ Non infected leaf
🟧 Infected but not detected leaf
🟥 Infected and detected leaf

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.

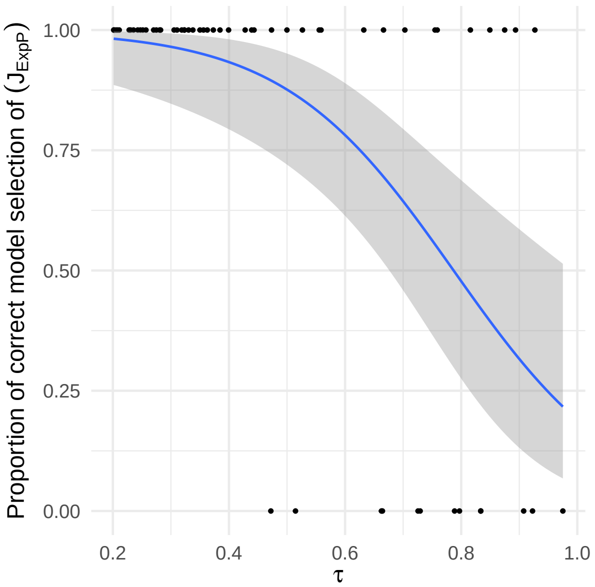

Figure 2: Logistic regression of the proportion of correct model selection of dispersal J_{ExpP} as a function of τ. Dots represent the values of τ used for the ~~50~~ 60 replicates of simulated dispersal model J_{ExpP}, and the estimated dispersal model (1 for a correct model selection of J_{ExpP} and 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%.

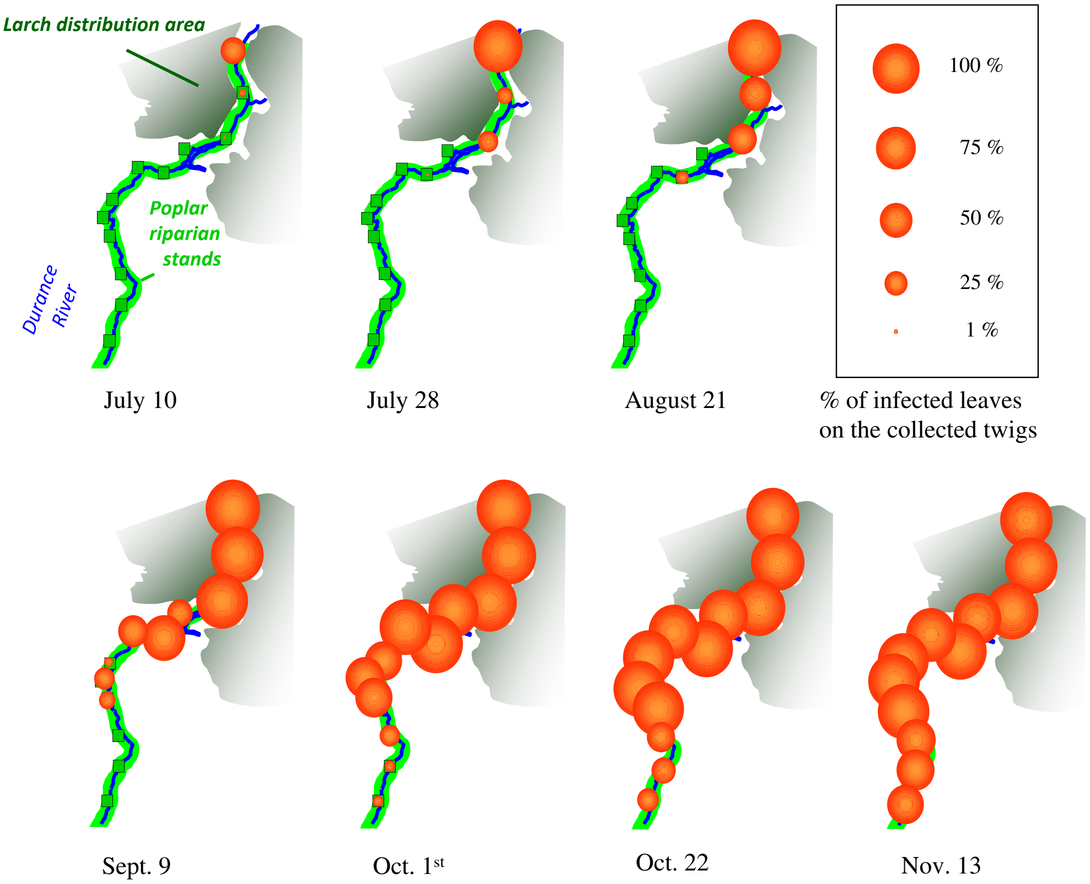

Figure 3: Poplar rust epidemic wave along the Durance River valley in 2008. The larch distribution area is represented in dark green, wild poplar riparian stands in pale green. The 12 study sites are represented by the green squares. Orange dots describe the evolution of the poplar rust epidemic through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

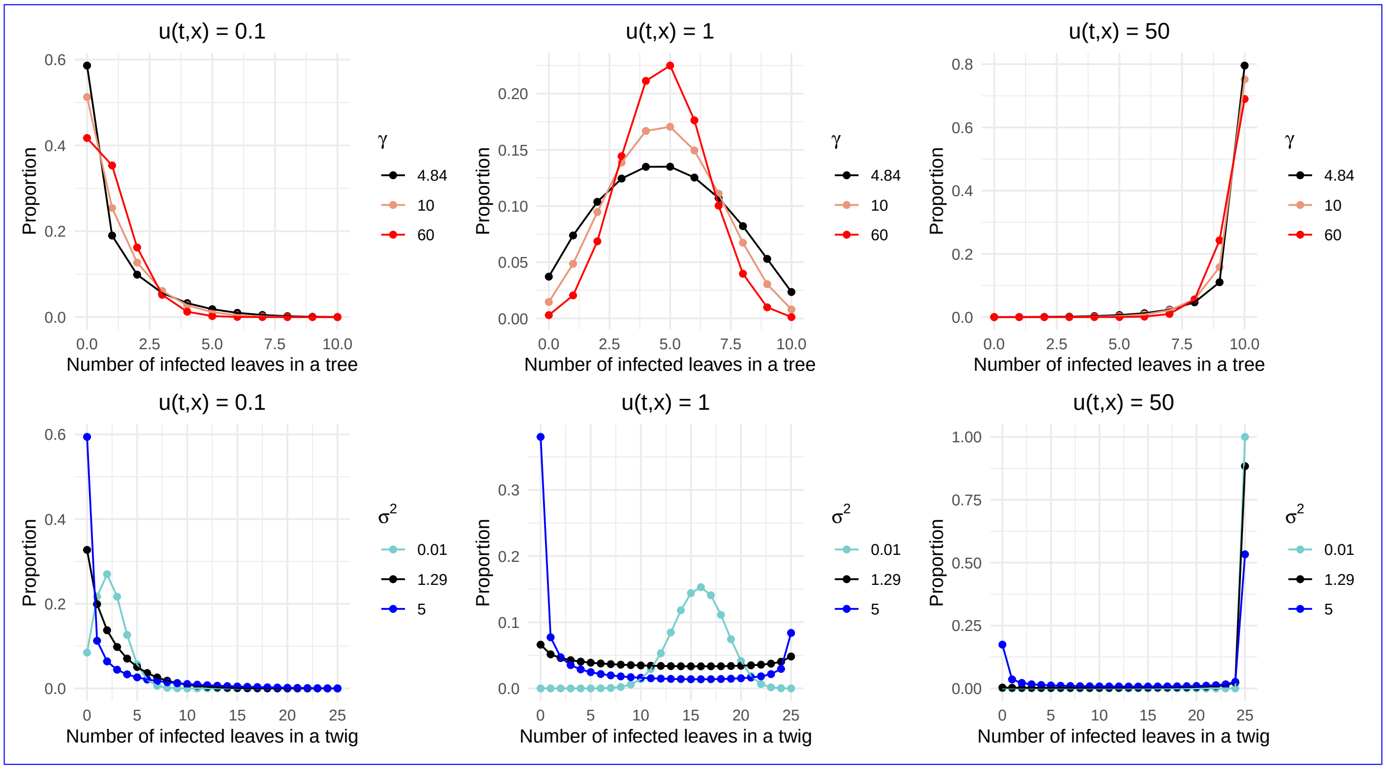

Figure 4: Distributions of the number of infected leaves in a tree and of the number of infected leaves in a twig, for increasing densities of infection $u(t,x)$, and contrasted levels of environmental heterogeneity σ^2 and γ. The number of infected leaves in a tree follows a Beta-Binomial distribution (Eq. (S12)) with ~~$\sigma^2 = 1.09$~~$\sigma^2 = 1.29$. Its density is plotted for three tree perceptions γ: ~~5.21~~ 4.82 (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. (S18)). Its density is plotted for three leaf suitabilities σ^2: ~~1.09~~ 1.29 (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).

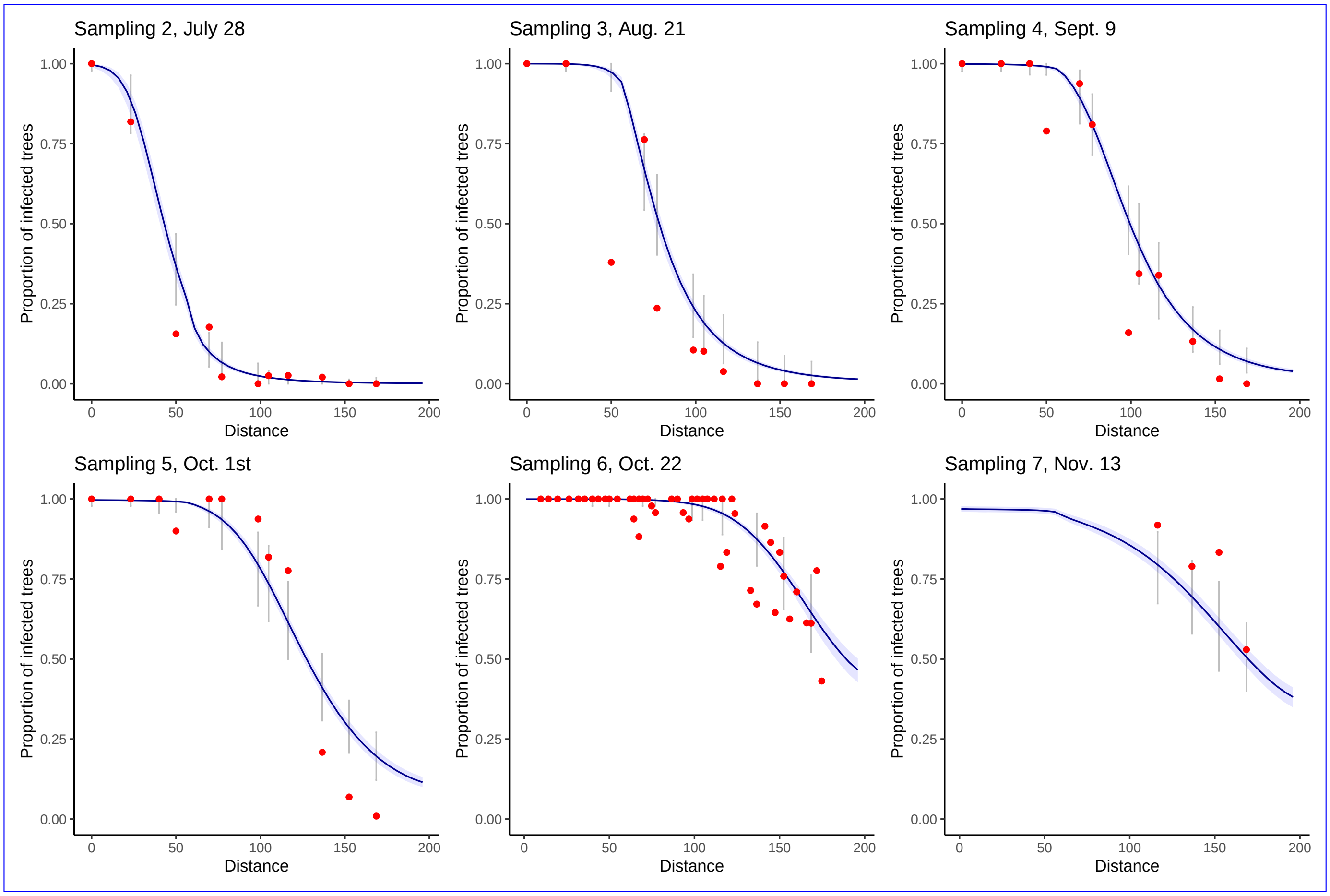

Figure 5: Model check under the selected dispersal model J_{ExpP}: Coverage rates for the raw sampling. Each sampling date is represented on a separate graph. Sampling 1 is not represented because it corresponds to the initial condition of the epidemics for all simulations. Blue areas correspond to the 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. The total coverage rate over all sampling dates is ~~0.75~~0.71.

Appendix to: A mechanistic-statistical approach to infer dispersal and demography from invasion dynamics, applied to a plant pathogen

Méline Saubin[1], Jérome Coville[2], Constance Xhaard[1,2,3], Pascal Frey[1], Samuel Soubeyrand[2], Fabien Halkett[1] and Frédéric 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

Corresponding author: Meline Saubin

Current adress: Populationsgenetik, Technische Universität München, Liesel-Beckmann-Str. 2, 85354 Freising, Germany

E-mail: meline.saubin@tum.de

S1 Numerical scheme

We use an implicit Euler scheme combined with a finite difference scheme (see Allaire, 2005 for details) to compute the solution $u(t,x)$ of the reaction-diffusion equation over $[-R,R] \times [0,T]$, with $2 \times R$ the length of the modelled environment, and T the duration of the modelled process. For the integro-differential equation, we use an explicit Euler scheme. More precisely, we perform a standard explicit Euler time discretisation of the equation:

$$\frac{\partial u}{\partial t}(t,x) \approx \frac{u(t+\delta,x)-u(t,x)}{\delta} \tag{S1}$$

that leads to:

$$\begin{aligned} u(t_{n+1},x) = u(t_n,x) + \delta \left(\int_{-R}^{R} J(x-y)[u(t_n,y)-u(t_n,x)]\, dy \right) \\ + \delta r(x) u(t_n,x) \left(1 - \frac{u(t_n,x)}{K} \right) \end{aligned} \tag{S2}$$

where $\{t_n = n\delta = nT/N : n = 0,\dots,N\}$ is a series of increasing times separated by $\delta = T/N > 0$, and N is the number of time steps in the series. For the space discretisation, we define a regular grid $\{x_i = -R + i\varepsilon = -R + 2Ri/I : i = 0,\dots,I\}$ with $I+1$ points separated by $\varepsilon = 2R/I > 0$. We make the following approximation for all x in $[-R,R]$:

$$u(t_n,x) \approx \sum_{i=0}^{I} u(t_n,x_i) \mathbb{1}_{[x_i,x_i+\varepsilon)}(x) \tag{S3}$$

where $x \mapsto \mathbb{1}_{[x_i,x_i+\varepsilon)}(x)$ is the indicator function that gives 1 if $x \in [x_i, x_i+\varepsilon)$, 0 otherwise. Based on this approximation, we only need to compute $u(t,x)$ at points x_i, $i = 0,\dots,I$. Plugging Approximation (S3) in the integral of Equation (S2) computed for $x = x_i$ yields:

$$\begin{aligned}
&\int_{-R}^{R} J(x_i - y)[u(t_n,y) - u(t_n,x_i)]\,dy \\
&\qquad \approx \int_{-R}^{R} J(x_i - y)\left[\left(\sum_{j=0}^{I} u(t_n,x_j)\mathbb{1}_{[x_j,x_j+\varepsilon)}(y)\right) - u(t_n,x_i)\right] dy \\
&\qquad = \left(\sum_{j=0}^{I} u(t_n,x_j)\int_{-R}^{R} J(x_i - y)\mathbb{1}_{[x_j,x_j+\varepsilon]}(y)\,dy\right) - \left(u(t_n,x_i)\int_{-R}^{R} J(x_i - y)dy\right) \\
&\qquad \approx \varepsilon\left(\sum_{j=0}^{I} u(t_n,x_j)J(x_i - x_j)\right) - \varepsilon\, u(t_n,x_i)\sum_{j=0}^{I} J(x_i - x_j)
\end{aligned} \tag{S4}$$

Let us define the matrix $\mathbf{J}^{in} := (J(x_i - x_j))_{0\leq i,j\leq I}$ whose element (i,j) is $\mathbf{J}^{in}_{ij} = J(x_i - x_j)$. We get the following numerical scheme:

$$\begin{aligned}
u(t_{n+1},x_i) = u(t_n,x_i) + \delta\varepsilon\left[\sum_{j=0}^{I}\mathbf{J}^{in}_{ij}u(t_n,x_j) - u(t_n,x_i)\left(\sum_{j=0}^{I}\mathbf{J}^{in}_{ij}\right)\right] \\
+ \delta r(x_i)u(t_n,x_i)\left[1 - \frac{u(t_n,x_i)}{K}\right]
\end{aligned} \tag{S5}$$

By defining the vectors $\mathbf{U}(t_n) = (u(t_n,x_i))_{0\leq i\leq I}$, $\mathbf{R} = (r(x_i))_{0\leq i\leq I}$ and $\mathbf{1} = (1)_{0\leq i\leq I}$, we have to solve the linear system:

$$\mathbf{U}(t_{n+1}) = \mathbf{U}(t_n) + \delta\varepsilon\left\{\mathbf{J}^{in}\mathbf{U}(t_n) - \mathbf{U}(t_n)\cdot(\mathbf{J}^{in}\mathbf{1})\right\} + \delta\{\mathbf{R}\cdot\mathbf{U}(t_n)\}\cdot\left\{\left(1 - \frac{\mathbf{U}(t_n)}{K}\right)\right\} \tag{S6}$$

where $\cdot$ is the element-wise multiplication operator.

S2 Distributions of the population measurements

S2.1 Term designations for the sampling units

In our biological application, a poplar leaf represents a habitat unit, a twig represents a group of habitat units, and a tree represents a habitat bloc. For clarity, we refer to leaves, twigs and trees in the following explanations. We call a sampling site a surveyed area along the valley, containing several hundreds of trees. Further adaptations of this model to other sampling units would only require adapting this initial vocabulary (Figure 1).

S2.2 Raw sampling

In the raw sampling, trees represent the sampling units, and B_{st} trees are observed in site s at time t. For each tree $b \in \{1,\dots,B_{st}\}$, we measure the presence/absence of the pathogen by monitoring an equivalent number of M leaves within b (see Appendix S3 below for the determination of M). A tree is infected if at least one pathogen lesion has been detected, in at least one leaf of the tree. The observation in site s at time t is the number Y_{st} of infected trees.

Now, let us derive the probabilistic law of the presence/absence of the pathogen in any tree b observed in site s at time t. In this paragraph, subscripts s, t, and b are generally omitted to avoid cumbersome notation. We first remind that the numbers of pathogen lesions $N_i(t)$ in the leaf $i \in \{1,\dots,M\}$ observed in tree b, given $R_i(t)$ and $u(t,x_s)$, are independent and Poisson distributed (see Eq. (2) in the main text):

$$N_i(t) \mid u(t,x_s), R_i(t) \underset{\text{indep.}}{\sim} \text{Poisson}(u(t,x_s)R_i(t)) \tag{S7}$$

In the raw sampling, M leaves are sampled at different locations on the tree (*i.e.* they belong to different groups, referred to as twigs), but further information about the twigs is not known. Thus, in the following, we take into account the twig structure without exploiting twig information. The leaves of a given twig g on tree b share at time t the same suitability $\mathcal{R}_g(t)$, which is unobserved and Gamma distributed like in Eq. (3) in the main text (for all leaves i in twig g, $R_i(t) = \mathcal{R}_g(t)$). Given the suitabilities $\{\mathcal{R}_g(t) : g = 1, \dots, G\}$ of twigs which compose tree b and given the absence of data about the twigs, $R_i(t)$ ($i \in \{1, \dots, M\}$) are independent and identically distributed under the discrete empirical probability distribution:

$$\hat{F}_G(r) = \frac{1}{G} \sum_{g=1}^{G} \mathbb{1}(r \leq \mathcal{R}_g(t)) \tag{S8}$$

where $\mathbb{1}(\cdot)$ is the indicator function. Therefore, $N_i(t)$ ($i \in \{1, \dots, M\}$) given $\{\mathcal{R}_g(t) : g = 1, \dots, G\}$ and $u(t, x_s)$ are independent and their probability distribution is, using Eqs. (S7)–(S8):

$$P[N_i(t) = n \mid u(t,x_s), \{\mathcal{R}_g(t) : g = 1, \dots, G\}] = \frac{1}{G} \sum_{g=1}^{G} \exp(-u(t,x_s)\mathcal{R}_g(t)) \frac{(-u(t,x_s)\mathcal{R}_g(t))^n}{n!} \frac{(u(t,x_s)\mathcal{R}_g(t))^n}{n!} \tag{S9}$$

The suitability $\mathcal{R}_g(t)$ being Gamma distributed with shape and scale parameters σ^{-2} and σ^2, respectively, the right-hand-side of Eq. (S9) is a Monte Carlo approximation of the integral:

$$\begin{aligned} &\int_{\mathbb{R}_+} \exp(-u(t,x_s)r) \frac{(-u(t,x_s)r)^n}{n!} \frac{(u(t,x_s)r)^n}{n!} \frac{1}{(\sigma^2)^{\sigma^{-2}}\Gamma(\sigma^{-2})} r^{\sigma^{-2}-1} e^{-r/\sigma^2} dr \\ &= \frac{\Gamma(n+\sigma^{-2})}{(n!)\Gamma(\sigma^{-2})} \left(1 - \frac{u(t,x_s)}{u(t,x_s)+\sigma^{-2}}\right)^{\sigma^{-2}} \left(\frac{u(t,x_s)}{u(t,x_s)+\sigma^{-2}}\right)^n \end{aligned} \tag{S10}$$

which coincides with the probability distribution of the Negative–Binomial law (*i.e.* the Gamma-

Poisson mixture distribution) given by Eq. (4) in the main text. The larger G, the more precise the approximation. Consequently, $N_i(t)$ ($i \in \{1,\ldots,M\}$) given $u(t,x_s)$ are asymptotically independent and distributed under the Negative–Binomial distribution given by Eq. (4) in the main text. Based on this approximation, the infections of leaves from tree b in site s at time t are asymptotically independent and distributed under Bernoulli distributions with success probability:

$$\begin{aligned} p_{st}^{\text{leaf}} &= P(N_i(t) > 0 \mid u(t,x_s)) \\ &= 1 - P(N_i(t) = 0 \mid u(t,x_s)) \\ &= 1 - (1 + u(t,x_s)\sigma^2)^{-1/\sigma^2} \end{aligned} \tag{S11}$$

The people who carried out the sampling observed a number M of leaves on tree b. Due to the particular configuration of the foliage of each tree, we assumed that the number Y_{stb}^{leaf} of infected leaves among the M leaves observed in tree b is approximately distributed under a Beta-Binomial distribution with mean Mp_{st}^{leaf} and tree perception parameter γ:

$$Y_{stb}^{\text{leaf}} \mid u(t,x_s) \sim_{approx.} \text{Beta-Binomial}(M, p_{st}^{\text{leaf}}, \gamma) \tag{S12}$$

Accordingly, the probability, as *perceived* by people in charge of the sampling, of leaf infection on the set of M leaves observed on a given tree, is distributed according to a Beta distribution. The Beta distribution is centred around the true probability of leaf infection p_{st}^{leaf} and allows *perceived* probability to vary from tree to tree depending on the tree perception parameter γ. It follows that the infection of tree b is approximately distributed under the Bernoulli distribution with success probability:

$$\begin{aligned}p_{st}^{\text{tree}} &= P(Y_{stb}^{\text{leaf}} > 0 \mid u(t,x_s)) \\ &= 1 - P(Y_{stb}^{\text{leaf}} = 0 \mid u(t,x_s)) \\ &= 1 - \frac{\text{Beta}[\gamma p_{st}^{\text{leaf}}, M + \gamma(1 - p_{st}^{\text{leaf}})]}{\text{Beta}[\gamma p_{st}^{\text{leaf}}, \gamma(1 - p_{st}^{\text{leaf}})]}\end{aligned} \tag{S13}$$

where p_{st}^{leaf} is given by S11 and Beta represents the beta function. It follows that the probability distribution functions of the number Y_{st}^{tree} of infected trees infected among the B_{st} trees observed satisfy, for all sampling sites s and sampling times t:

$$\begin{aligned}f_{st}^{\text{raw}}(y) &= P[Y_{st}^{\text{tree}} = y \mid u(t,x_s)] \\ &= f_{\text{Binomial}(B_{st}, p_{st}^{\text{tree}})}(y)\end{aligned} \tag{S14}$$

where $f_{Binomial}$ is the density of the Binomial distribution.

S2.3 Refined sampling

In the refined sampling, G_{st} twigs (*i.e.* groups of spatially connected leaves) are sampled in site s at time t. Here, the twig information (the number of twigs and the distribution of leaves on twigs) are known but the suitability $\mathcal{R}_g(t)$ of leaves in a twig g remains unobserved. The numbers of pathogen lesions $N_i(t)$ in the observed leaves $i \in \{1, \dots, M_{stg}\}$ of twig g given $\mathcal{R}_g(t)$ and $u(t,x_s)$ are independent and Poisson distributed:

$$N_i(t) \mid u(t,x_s), \mathcal{R}_g(t) \underset{\text{indep.}}{\sim} \text{Poisson}(u(t,x_s)\mathcal{R}_g(t)) \tag{S15}$$

Then, the numbers of infected leaves Y_{stg}^{leaf} (*i.e.* leaves with at least one pathogen lesion) given $\mathcal{R}_g(t)$ and $u(t,x_s)$ are independent and distributed under the following Binomial distributions:

$$Y_{stg}^{\text{leaf}} \mid u(t,x_s), \mathcal{R}_g(t) \underset{\text{indep.}}{\sim} \text{Binomial}(M_{stg}, 1-e^{-u(t,x_s)\mathcal{R}_g(t)}) \tag{S16}$$

In addition,

$$u(t,x_s)\mathcal{R}_g(t) \mid u(t,x_s) \underset{\text{indep.}}{\sim} \text{Gamma}(\sigma^{-2}, u(t,x_s)\sigma^2) \tag{S17}$$

Using Eqs. (S15)–(S17), Y_{stg}^{leaf} given $u(t,x_s)$ are independent and follow Gamma-Binomial mixture distributions:

$$\begin{aligned} f_{st}^{\text{ref}}(y) &= P[Y_{stg}^{\text{leaf}} = y \mid u(t,x_s)] \\ &= \int_0^\infty f_{\text{Binomial}(M_{stg},1-e^{-z})}(y) f_{\text{Gamma}(\sigma^{-2},u(t,x_s)\sigma^2)}(z)dz \end{aligned} \tag{S18}$$

where f_{Gamma} is the density of the Gamma distribution. Note that this Gamma-Binomial mixture distribution is an over-dispersed Binomial distribution like the Beta-Binomial distribution.

S3 Estimation of the number of leaves efficiently observed during tree scans

A problem inherent to the raw sampling design is that we do not know the number of leaves observed during the scan of the trees, contrary to the twig data for which we counted both the number of infected leaves and the total number of leaves carried by each observed twig. In other words, an inspected tree is a set of leaves of unknown size.

We assumes in Eq. (S12) that the number Y_{stb}^{leaf} of infected leaves among the M leaves observed in tree b is approximately distributed under a Beta-Binomial distribution with mean Mp_{st}^{leaf} and tree perception parameter γ. Parameter γ is however an unknown parameter. To overcome this parameter when calculating the average number of leaves observed per tree, we use the fact that on average the number of infected leaves is the same with a binomial distribution:

$$Y_{stb}^{\text{leaf}} \mid u(t,x_s) \sim_{approx.} \text{Binomial}(M, p_{st}^{\text{leaf}}) \tag{S19}$$

From this distribution, we obtain at each site s and date t the probability p_{st}^{tree} that a tree is infected as a function of both the probability p_{st}^{leaf} that a leaf is infected and the number M of leaves observed on a tree:

$$\begin{aligned} p_{st}^{\text{tree}} &= P(Y_{stb}^{\text{leaf}} > 0 \mid u(t, x_s)) \\ &= 1 - P(Y_{stb}^{\text{leaf}} = 0 \mid u(t, x_s)) \\ &= 1 - (1 - p_{st}^{\text{leaf}})^M \end{aligned} \tag{S20}$$

Thus, the number of leaves on a tree satisfies:

$$M = \frac{\log(1 - p_{st}^{\text{tree}})}{\log(1 - p_{st}^{\text{leaf}})} \tag{S21}$$

Let us use as approximations of p_{st}^{tree} the observed proportions q_{st}^{tree} of infected trees at sites s and dates t, and as approximations of p_{st}^{leaf} the observed proportions q_{st}^{leaf} of infected leaves (calculated from twig data). Then, an estimate $\hat{\lambda}_M$ of the mean number of leaves λ_M by tree is given by:

$$\hat{\lambda}_M = \text{round}\left(\frac{1}{N}\sum_{i=1}^{N} \frac{\log(1 - q_{st}^{\text{tree}})}{\log(1 - q_{st}^{\text{leaf}})}\right) \tag{S22}$$

with N the number of pairs (s, t) (*i.e.* sampling sites and dates) displaying both tree and twig data. Proportions of infection $q_{st}^{\text{tree}} = 1$ and $q_{st}^{\text{leaf}} = 1$ where approximated to $1 - 10^{-16}$ for numerical considerations. This procedure led to $\hat{\lambda}_M = 10$. This value may appear low. However, λ_M does not correspond to the actual mean number of leaves carried by an entire young tree but amounts to the mean number of leaves effectively inspected during tree scan, *i.e.* those observed as minutely as for the twig data in a limited time (see Eq. (S13)). It is important to note that for each tree the tree scan stops when an infected leaf is observed, or after 30 s of inspection. Therefore, the number of inspected leaves per tree can be very low in highly infected sites.

For the practical identifiability studies, we set $\lambda_M = 10$. For parameter inference on the real data set a different value of $(\hat{\lambda}_M)_t$ was estimated for each sampling date, from the observed proportions q_{st}^{tree} of infected trees and the observed proportions q_{st}^{leaf} of infected leaves at date t (Table S1).

Table S1: Estimated number of leaves effectively observed per tree for each sampling date t, $(\hat{\lambda}_M)_t$. The values of $(\hat{\lambda}_M)_t$ were used in the application on the real data set.

Date t	$(\hat{\lambda}_M)_t$
1	40
2	24
3	6
4	3
5	5
6	1

S4 Simulation details

Computations were performed with the R software environment (R Core Team, 2018). The vector of initial population densities $u(0,x)$ for x over $[-R,R]$ was estimated from the data of the first sampling date, by fitting a general model for analysis of dose-response data (package Drc on R, Ritz et al., 2015). This vector of initial population densities represented the initial condition of all simulations. We modelled $N = 1500$ time steps and $I = 400$ points in space. Because of the numerical scheme, with these parameters the reaction-diffusion dispersal model R.D. required an upper limit for parameter λ: we set $\lambda_{up} = 23$ for this model.

To fit our real case study, for all simulations we set $R = 100$ km, for a 200 km long river valley, and the epidemic was monitored over $T = 150$ days. We considered a shift in the environment topology at $d = 0.31\%$ of the valley, which corresponds to the delimitation observed in the Durance River valley with the Serre-Ponçon dam at 62 km downstream of the starting point of the epidemic. Therefore, for all simulations, the two growth rates r_{up} and r_{dw} apply to continuous segments of proportions d and $1-d$ of the monitored space, respectively.

S4.1 Practical parameter identifiability

Simulations were performed as follows in three steps.

Step 1 : Simulation of a realistic epidemic. Given a hypothetical dispersal model (J_{Exp}, J_{Gauss}, J_{ExpP} or R.D.), values in the parameter vector $\theta = (\theta_r, \theta_J, \gamma, \sigma^2)$ are independently and randomly drawn from dedicated distributions encompassing a large diversity of invading scenarios and specified in Table S1. We then simulate the corresponding epidemic along the 1D spatial domain $[-R,R]$. This epidemic is considered 'realistic' if a set of requirements on the observed proportion of infected trees $P_{s,t}$ on the farther downstream site ($s = R$) is met:

- $P_{R,30} < 0.1$ (the proportion of infected trees after one month is lower than 10%);
- $P_{R,75} < 0.5$ (the proportion of infected trees after two and a half months is lower than 50%);
- $P_{R,150} > 0.1$ (the proportion of infected trees after five months is higher than 10%);
- $P_{R,150} < 0.8$ (the proportion of infected trees after five months is lower than 80%).

Step 1 is complete once a candidate vector θ leads to an epidemic satisfying the four conditions described above (*i.e.* the simulation of θ and the epidemic is repeated while the four conditions are not satisfied). Thereafter, the vector finally retained in Step 1 is denoted θ_{true}.

Table S1: Marginal distributions used to randomly sample the model parameters included in $\theta = (\theta_r, \theta_J, \gamma, \sigma^2)$ before checking the requirements detailed in Step 1, with $\theta_r = (r_{\text{dw}}, \omega)$ and $\theta_J = (\lambda)$ or $\theta_J = (\lambda, \tau)$ depending on the model.

Parameter	Distribution	Interval
r_{dw}	Log-Uniform	[0.01, 0.5]
ω	Uniform	[-2, ~~3~~4]
λ	Log-Uniform	[0.2,5]
τ	Log-Uniform	[0.2,1]
γ	Log-Uniform	[2,20]
σ^2	Log-Uniform	[0.01, 15]

Step 2 : Simulation of the sampling process. We consider a sampling design similar to our real experiment with six sampling dates and 12 sampling locations regularly spread over 150 days and 200 km, respectively ($R = 100$ km). As for our real data, we increase the location density for the fifth date, with 45 locations instead of 12. For each date and location, the raw sampling consists in simulating the observed sanitary status of 10 leaves per tree from 100 trees, and the refined sampling consists in simulating the observed sanitary status of 25 spatially connected leaves from 20 twigs,

the simulations being performed given θ_{true}. The resulting data set is denoted $\mathcal{D}_{\text{true}}$.

Step 3 : Parameter estimation. We use the data $\mathcal{D}_{\text{true}}$ to estimate the model parameters by minimizing the logarithm of the likelihood function $L(\theta)$. In our case, preliminary tests revealed that classical optimisation algorithms were not accurate enough to provide satisfactory rates of convergence due to local optimum problems. Thus, we adopt a hybrid strategy combining first a Nelder-Mead algorithm (improving global search ability) and then a Nlminb algorithm (for its high computational efficiency). Specifically, we proceed in three substeps described below, the crucial stage consisting in finding initial values that give a satisfactory rate of convergence.

Step 3.1 : Using Step 1, we generate 500 vectors θ_{init}. Note that this step was only performed once for all the estimations performed in this article. We provide in Figure S1 a comparison of the initial distribution of parameters as stated in Table S1, and of the distribution of parameters in the vector θ_{init}, *i.e.* leading to "realistic" epidemics.

Step 3.2 : The corresponding 500 likelihood values $L(\theta_{\text{init}})$ are calculated given $\mathcal{D}_{\text{true}}$. Then, the 20 vectors θ_{init} corresponding to the 20 largest likelihood values are used as initial values for 50 steps of a NELDER-MEAD optimisation routine (`R` function `optim`), resulting in 20 updated initial parameter vectors θ_{init2} depending on $\mathcal{D}_{\text{true}}$. The new initial vectors θ_{init2} that do not satisfy lower bounds θ_{low} and upper bounds θ_{up} are excluded. We used $\theta_{\text{low}} = (r_{\text{dw}} = 0.001, \omega = -7, \lambda = 0.02, \tau = 0.02, \gamma = 1.05, \sigma^2 = 10^{-7})$ and ~~$\theta_{\text{up}} = (r_{\text{dw}} = 0.5, \omega = 3, \lambda = 10, \tau = 1, \gamma = 30, \sigma^2 = 20)$~~$\theta_{\text{up}} = (r_{\text{dw}} = 0.5, \omega = 4, \lambda = 10, \tau$ with $\lambda = 23$ in θ_{up} instead of 10 for the R.D. model. The validity intervals defined by θ_{low} and θ_{up} encompass the intervals used to simulate θ (see Table S1). The likelihood values of the n_{init} remaining vectors $L(\theta_{\text{init2}})$ are calculated (given $\mathcal{D}_{\text{true}}$) and ranked in descending order.

Step 3.3 : θ is then estimated using the NLMINB optimisation routine with lower and upper bounds θ_{low} and θ_{up}, respectively. The initial parameter values are set to the first vector θ_{init2} as ordered in the previous step. The estimated parameter values, say θ_{estim}, are accepted if the `nlminb` function in `R` delivered a successful convergence diagnostic (with tunning parameters `rel.tol`=5.10^{-5} and `iter.max`=3000). If not, the second vector θ_{init2} is used, and so on until reaching convergence or testing the n_{init} initial vector's values selected at step 3.2. In the latter case, a convergence failure is obtained. Overall, this algorithm allows to obtain high rates of convergence.

These three steps were reiterated until deriving the estimation of ~~$n = 100$~~ $n = 160$ realistic epidemics for each dispersal model. Checking for practical identifiability of parameters basically relies on plotting for each dispersal model the cloud of points between θ_{true} and θ_{estim} (Figures S2, S3, S4, S5) and computing the corresponding correlations. Among all simulations performed, the proportions of convergence were ~~0.91, 0.95, 0.93, and 0.90~~ 0.98 for dispersal J_{Exp}, J_{Gauss}, ~~J_{ExpP},~~ and R.D. ~~, respectively~~and up to 0.99 for J_{ExpP}. A simulation converged when the convergence diagnostic of the algorithm indicated a convergence, and when all parameters were estimated inside intervals defined by θ_{low} and θ_{up}. In the small number of simulations where the value of λ_{estim} proposed by the optimisation algorithm was higher than 23 (which is the upper limit of our numerical scheme, Appendix S1), the simulation was still considered convergent with $\lambda_{\text{estim}} = 23$. This configuration can occur in particular when trying to fit dispersal R.D. on datasets simulated according to J_{ExpP}.

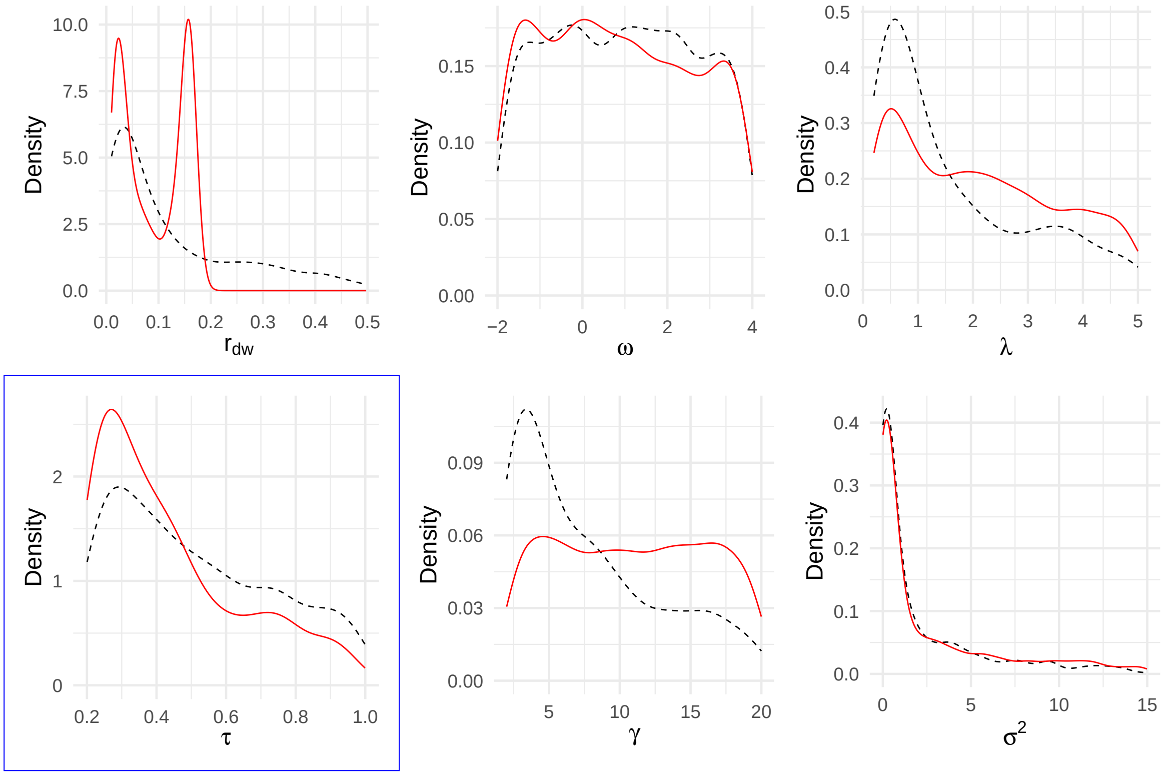

Figure S1: Distributions of parameters, before (in dotted black) and after (in red) retaining only parameters values leading to "realistic" epidemics. Dotted black distributions correspond to distributions given by Table S1. Red line distributions correspond to the distribution of parameters in θ_{init}. We represent here the distribution of "realistic" epidemics from the four hypothetical dispersal models (J_{Exp}, J_{Gauss}, J_{ExpP} and R.D.) for parameters r_{dw}, ω, γ and σ^2, for J_{ExpP} for parameter τ, and for J_{Exp}, J_{Gauss} and J_{ExpP} for parameter λ.

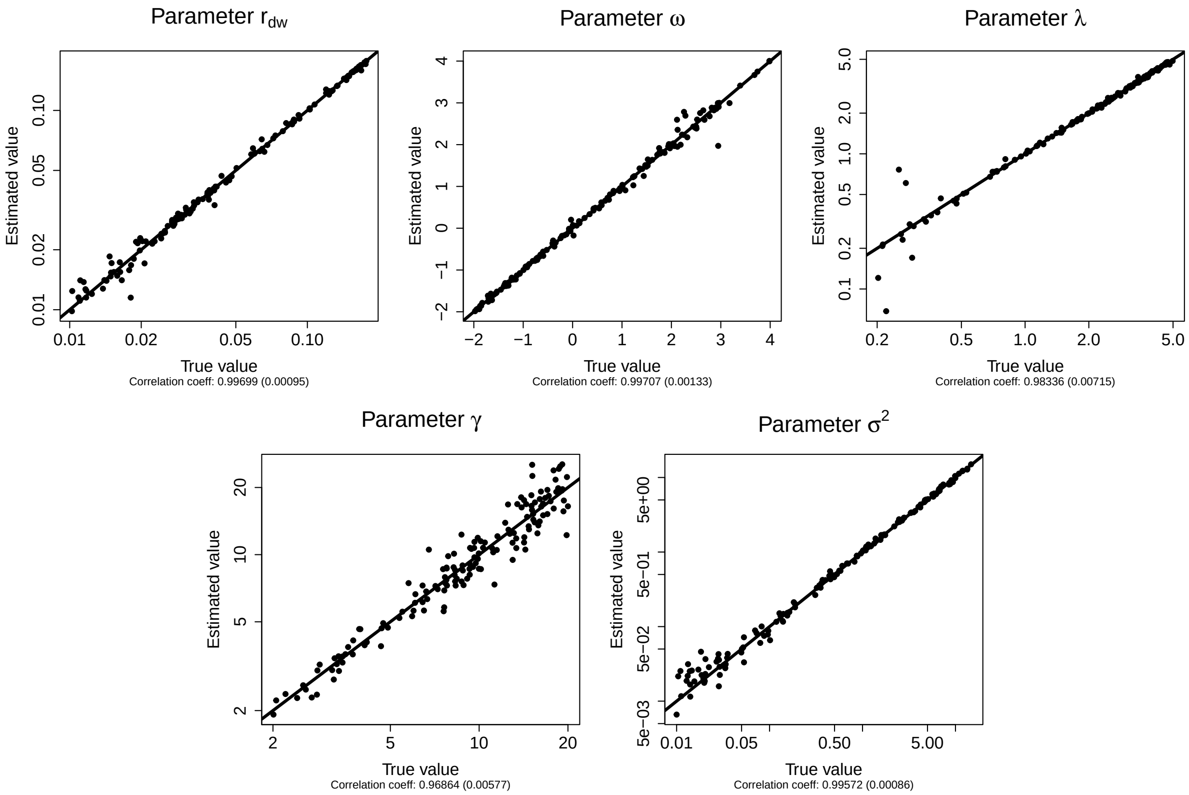

Figure S2: Practical parameter identifiability for the dispersal model J_{Exp}. Each point represents the parameter estimation ('Estimated' value) depending on the real parameter ('True' value). Each graph regroups the results of ~~100~~ 160 replicates. Straight lines correspond to the first bisector.

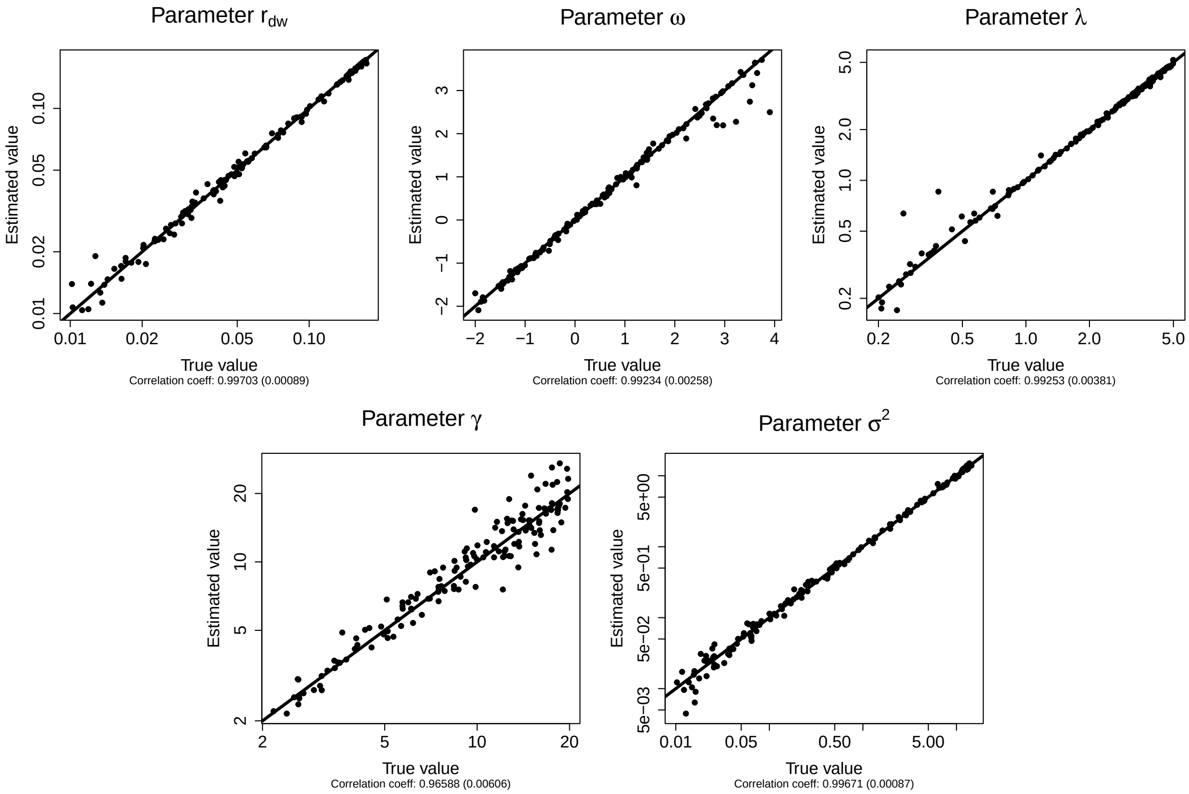

Figure S3: Practical parameter identifiability for the dispersal model J_{Gauss}. Each point represents the parameter estimation (‘Estimated’ value) depending on the real parameter (‘True’ value). Each graph regroups the results of ~~100~~ 160 replicates. Straight lines correspond to the first bisector.

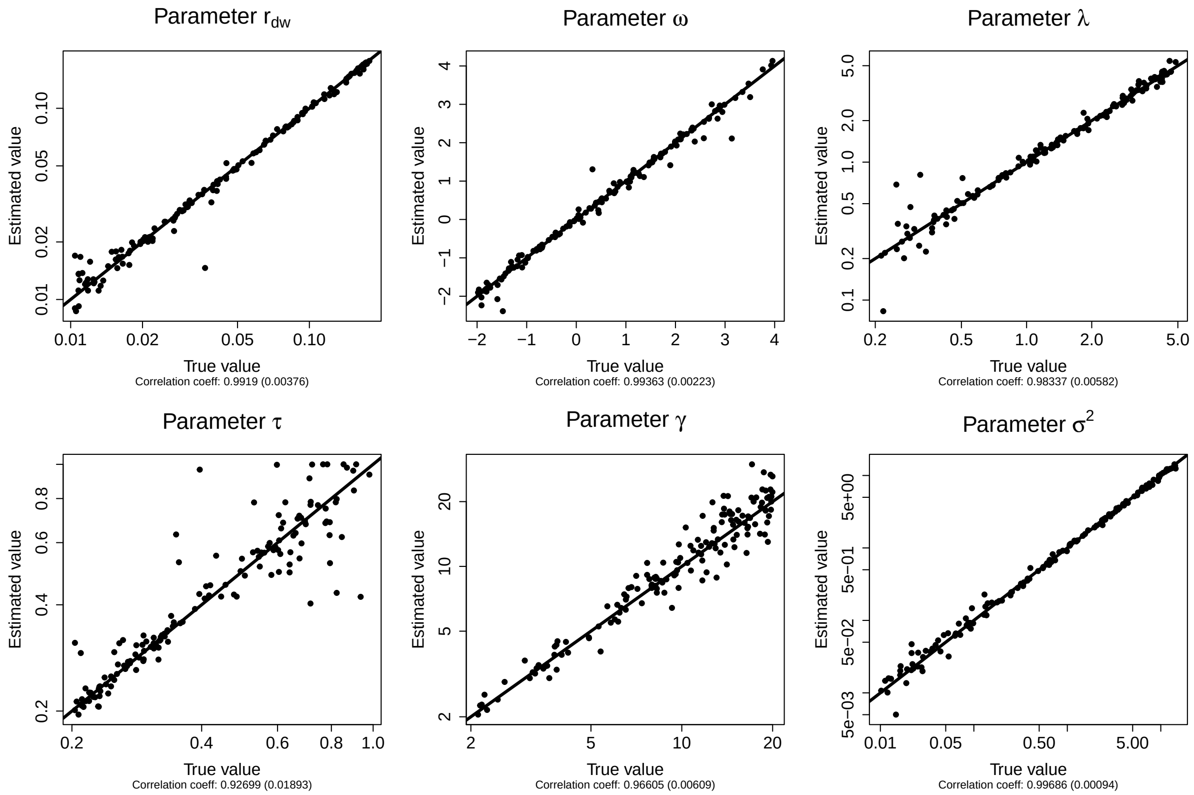

Figure S4: Practical parameter identifiability for the dispersal model J_{ExpP}. Each point represents the parameter estimation (‘Estimated’ value) depending on the real parameter (‘True’ value). Each graph regroups the results of ~~100~~ 160 replicates. Straight lines correspond to the first bisector.

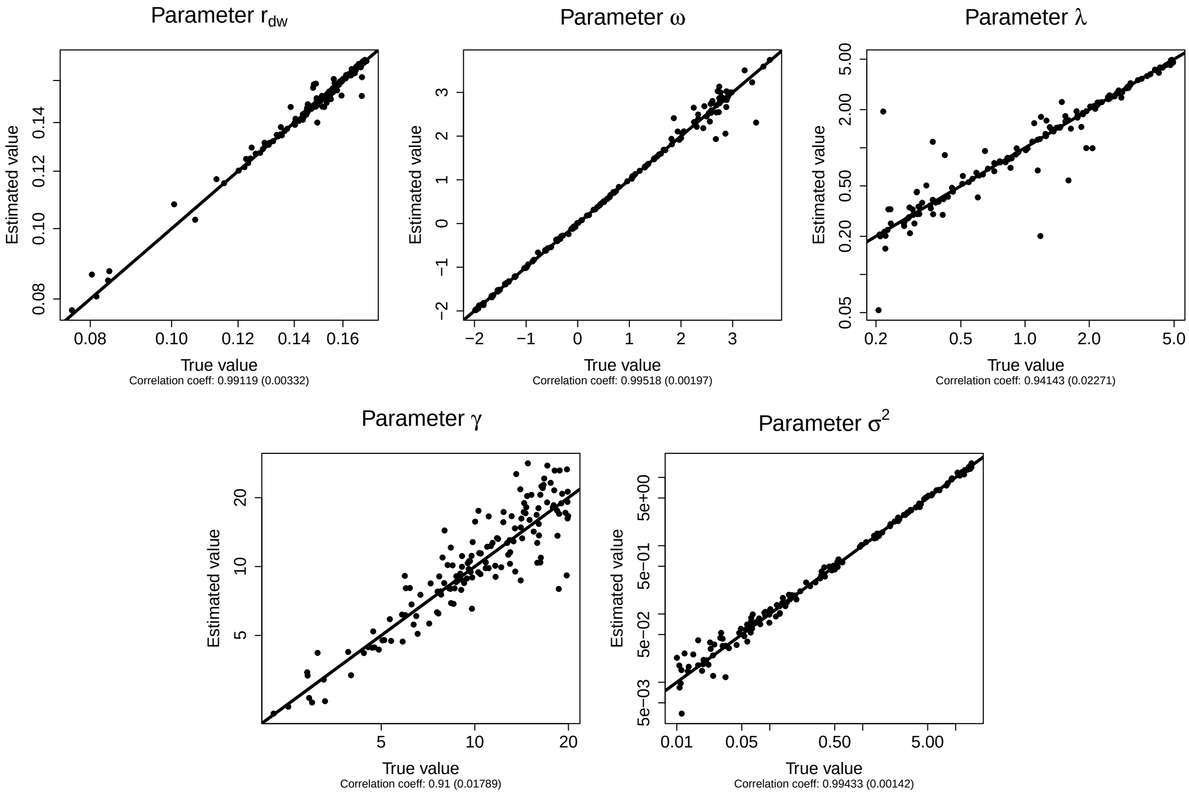

Figure S5: Practical parameter identifiability for the dispersal model R.D. Each point represents the parameter estimation ('Estimated' value) depending on the real parameter ('True' value). Each graph regroups the results of ~~100~~ 160 replicates. Straight lines correspond to the first bisector.

S4.2 Model selection

Model practical identifiability was carried out in a similar way than parameter practical identifiability (Appendix S4.1), except that we fitted to each data set the true model (as previously) but also the three other models corresponding to the alternative hypotheses on the dispersal process. Models were compared using AIC (Akaike Information Criteria) to select the best data-supported model. AIC were assessed as $2k - 2ln(L)$ where k is the number of parameters of the model considered and L is the maximized value of the likelihood function. To gain more insights into the confidence level in model selection, we also calculated for each data set the difference between the AIC of the model selected and the AIC of the second-best model according to the two possible issues of the selection procedure: (i) when the model selection procedure was successful (*i.e.* the selected model was the true model) and (ii) when the model selection procedure was incorrect (*i.e.* the true model was not selected). The mean of these values were reported as $d\text{AIC}_{\text{true}}$ and $d\text{AIC}_{\text{wrong}}$ in Table 2. The steps were reiterated until the estimation of $n = 50$ realistic epidemics for each dispersal model.

S4.3 Parameter inference on the real data set

The model selection procedure was applied to the real data set by fitting four dispersal process hypotheses (J_{Exp}, J_{Gauss}, J_{ExpP} and R.D.). The same optimisation routines described in Appendix S4.1 were performed from five initial parameter values selected as in Step 3.2 (Appendix S4.1). The selected model corresponds to hypothesis J_{ExpP}. For parameter estimations, we used the `mle2` function from the `R` package `bbmle`, with method NELDER-MEAD and optimizer NLMINB, to obtain maximum likelihood estimates of the vector of parameters $\hat{\theta}$ and of its matrix of variance-covariance $\hat{\Sigma}$. We used as initial parameter values the vector of parameters θ giving the lowest AIC value in the previous model selection procedure. Confidence intervals were derived from 1,000 random draws from the multivariate normal distribution with parameters $\hat{\theta}$ and $\hat{\sum}$. The 95% confidence intervals of each parameter is obtained using the quantiles 2.5% and 97.5% (Table 4).

S4.4 Model check

The model was checked by assessing the coverage rate of the data from the 95%-prediction intervals. The coverage rate was estimated as the proportion of observed data from the raw sampling within the prediction intervals (Figure 5).

Data from the raw sampling represent 97 counts Y_{st} of infected trees at sites $s \in \{1,\ldots,S\}$ (with $S = 12$ or $S = 45$ depending on the sampling date) and times $t \in \{1,\ldots,6\}$. Let us recall that, as stated in Appendix S2, Y_{st} follows a combination of Poisson and Beta-Binomial distributions whose parameters depend on the known mean value $(\lambda_m)_t$ and the unknown $u(t,x_s)$, γ and σ^2, and that $u(t,x_s)$ is a deterministic function of dynamical parameters r, λ and τ.

Prediction intervals were calculated at each date and each site with a two-step procedure:

Step 1 A confidence interval was obtained from 1000 random draws from the multivariate normal distribution with $\hat{\theta}$ and $\hat{\sum}$.

Step 2 The mean proportions of infected trees were calculated at each date and site date from each random draw of parameters obtained from Step 1. A prediction interval was obtained from these parameters given the probabilities of infection, with 1,000 random draws in the observation laws.

Model checks were performed for each dispersal kernel model, and not only the selected model J_{ExpP}, to ensure that the coverage rates were higher with the selected model (Figure 5 for the selected dispersal model J_{ExpP}, and Figures S6 and S7 for dispersal models J_{Exp} and J_{Gauss}, respectively). The model check was not performed for dispersal model R.D. because the estimated dispersal distance λ_{estim} reached the upper limit of our numerical scheme $\lambda_{\text{up}} = 23$ and did not allow to calculate the confidence intervals.

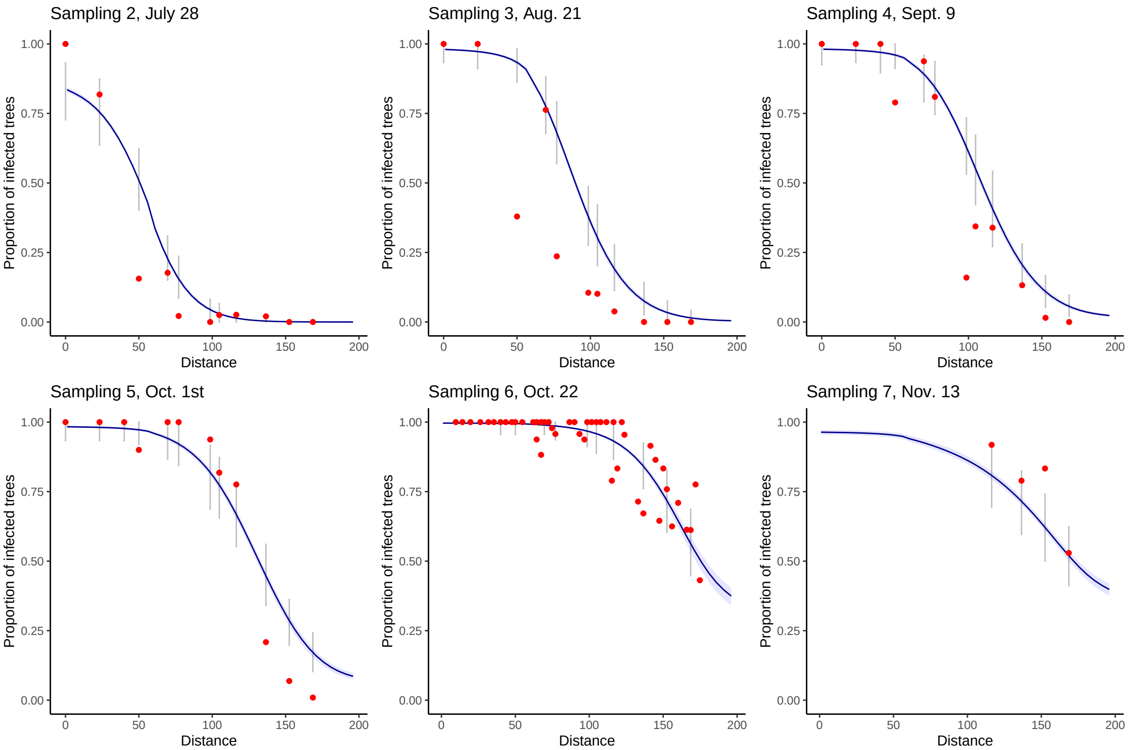

Figure S6: Model check under the dispersal model J_{Exp}: Coverage rates for the raw sampling. Each sampling date is represented on a separate graph. Sampling 1 is not represented because it corresponds to the initial condition of the epidemics for all simulations. Blue areas correspond to the pointwise 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. The total coverage rate over all sampling dates is ~~0.69~~0.68.

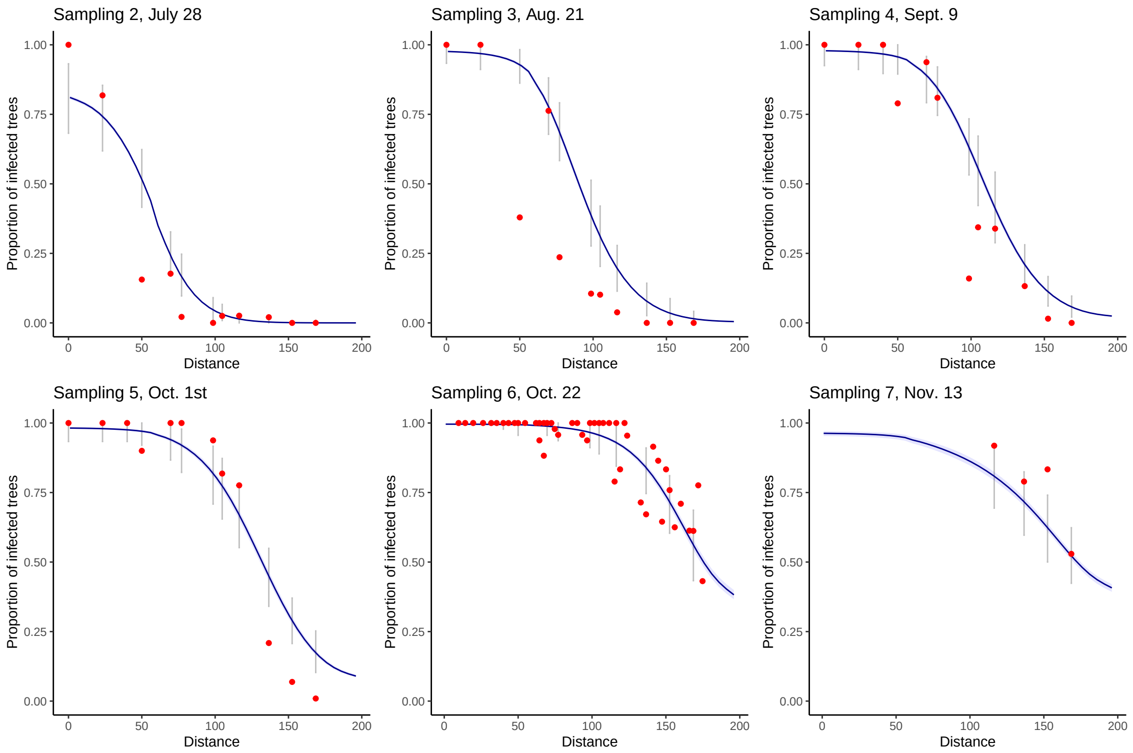

Figure S7: Model check under the dispersal model J_{Gauss}: Coverage rates for the raw sampling. Each sampling date is represented on a separate graph. Sampling 1 is not represented because it corresponds to the initial condition of the epidemics for all simulations. Blue areas correspond to the pointwise 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. The total coverage rate over all sampling dates is 0.67.

S4.5 Sampling densification

As in Appendix S4.2, numerical simulations were run to disentangle the true dispersal process from alternative dispersal processes, with densification of time and site for the raw and the refined sampling. Simulations were run with 21 sampling dates instead of 6, which amounts to one sampling every week. The number of sampling sites was set to 45 for all sampling dates. The steps described in Appendix S4.1 and S4.2 were reiterated until the estimation of $n = 50$ realistic epidemics for each dispersal model.

Table S2: Efficiency of model selection for the densification of time samples (21 instead of 6) and the site sampled (45 instead of 12). 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}	**~~0.72~~0.64**	~~0.06~~ 0.18	~~0.16~~ 0.12	0.06	~~3.23~~ 5.60	~~1.05~~2.02
J_{Gauss}	~~0.22~~ 0.14	**~~0.60~~0.8**	~~0.04~~ 0	~~0.14~~ 0.06	~~7.33~~ 9.93	~~1.67~~1.51
J_{ExpP}	~~0.12~~ 0.1	~~0.06~~ 0.02	**~~0.82~~0.86**	~~0~~ 0.02	**~~1788.56~~2228.45**	~~2.72~~1.81
R.D.	~~0.1~~ 0.12	~~0.28~~ 0.16	~~0.02~~ 0	**~~0.60~~0.72**	~~27.01~~ 32.97	~~0.94~~1.04

S5 Carrying capacity of poplar leaves

We measured the area of 10 wild poplar leaves (*Populus nigra*) and obtained a mean leaf area of 870 mm^2. We consider that poplar rust can not infect the leaf veins and edges, which represent approximately 15% of the leaf area. This leads to a net leaf area accessible to the pathogen of 740 mm^2. The size of a poplar rust lesion ranges from 0.2 mm^2 to 0.8 mm^2 (Maupetit et al., 2018). The lesions cannot fuse and are surrounded by living host tissue. We thus consider a lesion occupies a total area of 1 mm^2. This leads to a maximum of 740 lesions per leaf on average. To respect this order of magnitude, we consider in this analysis that the carrying capacity of a poplar leaf is 750 poplar rust lesions.

References

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

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.

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

Ritz, C., Baty, F., Streibig, J. C., and Gerhard, D. (2015). Dose-Response analysis using R. *PLoS One*, 10(12).