PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape

David R. J. Pleydell[1,2], Samuel Soubeyrand[3], Sylvie Dallot[1], Gérard Labonne[1], Joël Chadœuf[3], Emmanuel Jacquot[1], Gaël Thébaud[1]*

1 BGPI, INRA, Montpellier SupAgro, Univ. Montpellier, Cirad, TA A-54/K, Campus de Baillarguet, 34398, Montpellier cedex 5, France, **2** ASTRE, INRA, CIRAD, Univ. Montpellier, Montpellier, France, **3** BioSP, INRA, 84914, Avignon, France

* gael.thebaud@inra.fr

Citation: Pleydell DRJ, Soubeyrand S, Dallot S, Labonne G, Chadœuf J, Jacquot E, et al. (2018) Estimation of the dispersal distances of an aphid-borne virus in a patchy landscape. PLoS Comput Biol 14(4): e1006085. https://doi.org/10.1371/journal.pcbi.1006085

Editor: Samuel Alizon, CNRS, FRANCE

Received: May 5, 2017

Accepted: March 3, 2018

Published: April 30, 2018

Data Availability Statement: The sharka surveillance dataset used in this study contains confidential information and is owned by the French Plant Health Services. Applications to access this dataset should be sent to Christine Colas (SRAL/DRAAF; christine.colas@agriculture.gouv.fr).

Funding: This work was supported by: European Union (SharCo, FP7 204429), Département Santé des Plantes et Environnement, Institut National de la Recherche Agronomique, and FranceAgriMer. The funders had no role in study design, data

Abstract

Characterising the spatio-temporal dynamics of pathogens *in natura* is key to ensuring their efficient prevention and control. However, it is notoriously difficult to estimate dispersal parameters at scales that are relevant to real epidemics. Epidemiological surveys can provide informative data, but parameter estimation can be hampered when the timing of the epidemiological events is uncertain, and in the presence of interactions between disease spread, surveillance, and control. Further complications arise from imperfect detection of disease and from the huge number of data on individual hosts arising from landscape-level surveys. Here, we present a Bayesian framework that overcomes these barriers by integrating over associated uncertainties in a model explicitly combining the processes of disease dispersal, surveillance and control. Using a novel computationally efficient approach to account for patch geometry, we demonstrate that disease dispersal distances can be estimated accurately in a patchy (i.e. fragmented) landscape when disease control is ongoing. Applying this model to data for an aphid-borne virus (*Plum pox virus*) surveyed for 15 years in 605 orchards, we obtain the first estimate of the distribution of flight distances of infectious aphids at the landscape scale. About 50% of aphid flights terminate beyond 90 m, which implies that most infectious aphids leaving a tree land outside the bounds of a 1-ha orchard. Moreover, long-distance flights are not rare–10% of flights exceed 1 km. By their impact on our quantitative understanding of winged aphid dispersal, these results can inform the design of management strategies for plant viruses, which are mainly aphid-borne.

Author summary

In spatial epidemiology, dispersal kernels quantify how the probability of pathogen dissemination varies with distance from an infection source. Spatial models of pathogen spread are sensitive to kernel parameters; yet these parameters have rarely been estimated using field data gathered at relevant scales. Robust estimation is rendered difficult by practical constraints limiting the number of surveyed individuals, and uncertainties concerning their disease status. Here, we present a framework that overcomes these barriers to permit inference for a between-patch transmission model. Extensive simulations show

collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

that dispersal kernels can be estimated from epidemiological surveillance data. When applied to such data collected from more than 600 orchards during 15 years of a plant virus epidemic our approach enables the estimation of the dispersal kernel of infectious winged aphids. This kernel is long-tailed, as 50% of infectious aphids leaving a tree terminate their infectious flight beyond 90 m whilst 10% fly beyond 1 km. This first estimate of flight distances at the landscape scale for aphids–a group of vectors transmitting numerous viruses–is crucial for the science-based design of control strategies targeting plant virus epidemics.

Introduction

Infectious diseases of humans, animals and plants severely impact the world's health and economy. To gain knowledge on disease dynamics, powerful mathematical models have been developed [1–3]. However, for predicting the relative efficacies of competing control strategies across realistic heterogeneous landscapes, spatially-explicit *in silico* simulation models provide the main avenue [2]. The dispersal parameters of such models critically affect the predicted spatio-temporal dynamics of the disease, and thus the predicted outcome of potential control strategies [4]. Obtaining reliable estimates for these parameters is therefore a fundamental issue in epidemiology [5–7]. Models frequently employ dispersal kernels to represent how the probability of dispersal events diminishes as a function of distance, and simulation studies have proven that dispersal parameters can be identified in idealised scenarios [5]. Indeed, this has been achieved for simple models or small-scale datasets [8–13]. Recent advances in Bayesian methods and computing power have enabled fitting more realistic models to larger-scale surveillance data [6, 14–19]. However, most dispersal kernels are still unknown. Indeed, estimation gets more complex when graduating from idealised toy problems to reconstructing the spatio-temporal dynamics of real epidemics. The first issue is the mismatch between the spatio-temporal coordinates of the epidemic, sampling and model [20]. For example, the timing of key events (e.g. when a susceptible individual becomes infected) is often censored (i.e. known only within certain bounds), and failure to account for this can bias estimates. Moreover, the challenge of inference is increased by uncertainty arising from missing observations [21, 22] or imperfect sensitivity of disease detection [23, 24]. Further difficulties arise when surveillance data are aggregated at the patch scale because a landscape comprising patches of various shapes or sizes often cannot be summarized by patch centroids without biasing connectivity estimates. All these issues require appropriate correction measures to avoid biased inference and prediction [25].

In the case of aerial vector- or wind-borne diseases, dispersal kernels critically depend on the flight properties of the vectors or infectious propagules [26]. When the probability of dispersal decreases more slowly than an exponential distribution, kernels are termed "long-tailed" and lead to non-negligible long-distance flights [27]. Such events are an important component of disease epidemiological–and evolutionary–dynamics and call for kernel estimation at the landscape scale [28]. However, among plant diseases, there are few available kernel estimates. The dispersal kernel of black Sigatoka (a fungal disease of banana) has been estimated experimentally up to 1 km from a point source, based on the direct observation of spore-induced lesions [29]. This is the only available direct estimate at this scale for the dispersal kernel of a plant disease, which reflects the extreme practical difficulties of such field studies and highlights the critical need for developing *in silico* solutions. A promising way forward is to infer dispersal parameters indirectly, i.e. from spatio-temporal patterns observed in epidemiological data [5] whilst

accounting for the added complexity (outlined above) of observational studies. This approach has been used to infer the dispersal kernels of the wind-dispersed plantain fungus *Podosphaera plantaginis* [15], the fungus *Leptosphaeria maculans* affecting oilseed rape and dispersed both by wind and wind-driven rain [30], and two pathogens transmitted only by wind-driven rain: the oomycete *Phytophthora ramorum* that is responsible for sudden oak death [16], and the bacterium *Xanthomonas axonopodis* that causes Citrus canker [17]. A dispersal kernel has been estimated for two other Citrus diseases: Bahia bark scaling of Citrus, a disease with an elusive etiology [13], and Huanglongbing, which is caused by bacteria from the '*Candidatus* Liberibacter' genus and transmitted by psyllids [18]. To date, this is the only vector-borne plant disease for which the dispersal kernel is documented. Although aphids are responsible for transmitting almost 40% of more than 700 plant viruses [31] and impose large economic burdens, their dispersal remains ill-characterized at the landscape scale [32, 33]. For a vast number of aphid-borne diseases, this lack of basic knowledge affects science-based control strategies by undermining the reliability of quantitative risk assessment and predictive epidemiological models.

Most aphid-borne viruses belong to the *Potyvirus* genus and are transmitted in a non-persistent manner, i.e. by winged aphids that acquire and transmit the virus immediately while probing on various plants in search of a suitable host species [31]. Potyviruses are transmitted by a wide range of aphid species, and aphid infectivity is lost after the first probes. For these reasons, estimating the natural dispersal kernel of a potyvirus provides an indirect way of estimating the dispersal kernel of infectious winged aphids. *Plum pox virus* (PPV) is a potyvirus that is listed as one of the 10 most important plant viruses [34]. This virus is the causal agent of sharka, a quarantine disease affecting trees of the *Prunus* genus (i.e. mainly peach, apricot and plum), reducing fruit yield, quality (modified sugar content and texture) and visual appeal (due to deformations and discolouration) [33]. Sharka is a worldwide plague that has infected over 50 countries in Europe, Asia, America and Africa [33], inflicting estimated economic losses of 10 billion Euros over 30 years [35]. The transfer of infected (possibly symptomless) plant material can disseminate PPV over long distances [35], and the natural spread of the disease is ensured by more than 20 aphid species [36]. Virus-infected trees cannot be cured, and insecticides do not act fast enough to prevent the spread of the virus by non-colonising aphids [31, 37]. In addition, resistant or tolerant peach and apricot varieties are too scarce to provide a short-term alternative to cultivated varieties. However, aphid-mediated transmission can be reduced by removing infected trees as soon as they are detected. As a result, various countries have implemented PPV eradication or control strategies based on regular surveys and removal of trees or orchards when PPV is detected [33, 35, 38]. Given the cost of surveillance, tree removal and compensation, these strategies should benefit from model-assisted optimisation, which requires estimating the aphid dispersal kernel.

In this context, the aims of this study are: (i) to develop a Bayesian inference framework for estimating, from surveillance data, the parameters of a spatially-explicit epidemiological model that accounts for patch geometry and for interactions between disease spread, surveillance and control, (ii) to assess through simulations the accuracy and precision of the dispersal parameters estimated under various epidemic scenarios, and (iii) to apply our method to 15 years of geo-referenced surveillance data collected during an epidemic of *Plum pox virus* in order to estimate the dispersal kernel of the aphid vectors.

Materials and methods

Surveillance database

In the early 1990's, an outbreak of the M strain of PPV was detected in peach/nectarine patches (orchards) in southern France [39]. The plant health services implemented a control strategy

based on disease surveillance and removal of symptomatic trees. This process involved the routine collection of patch-level data comprising the observed number of new cases (trees with PPV-typical discolouration symptoms on flowers and leaves) and the corresponding inspection dates, as well as patch attributes (location, planting and removal years, planting density, etc.). We aggregated the information about a 5.6×4.8 km production area over surveillance years 1992-2006 into a unique georeferenced database, with patch boundary coordinates obtained from digitised aerial photographs. With 4820 inspections over 15 years in 553 patches (mean area: 0.95 ha; 52 orchards were replanted in these patches during that period), this database is a precious resource for inference on aphid-mediated viral dispersal in patchy (i.e. fragmented) landscapes. Moreover, to account for seasonal variation in the number of flying aphids, we used in our model the average (over 17 years) weekly number of flying aphids collected from a 12-m-high Agraphid suction tower located within the bio-geographical region of the study area.

Modelling framework

Our model has a compartmental Susceptible-Exposed-Infectious-Removed (*SEIR*) structure that aims to reduce bias in parameter estimates by accounting for irregular patch geometry, detection-dependent removal, imperfect detection sensitivity, interval censoring of between-compartment transition times, missing data and parameter uncertainty. We address these challenges by: (i) integrating a mixture of exponential dispersal kernels over source and receiver patches to compute between-patch connectivity; (ii) splitting the infectious state *I* into hidden (*H*) and detected (*D*) sub-states (Fig 1); (iii) integrating over uncertainty in the times of transition between compartments; (iv) using Bayesian data augmentation and inference. Two versions of our discrete-time spatio-temporal *SEHDR* model–one for stochastic simulations

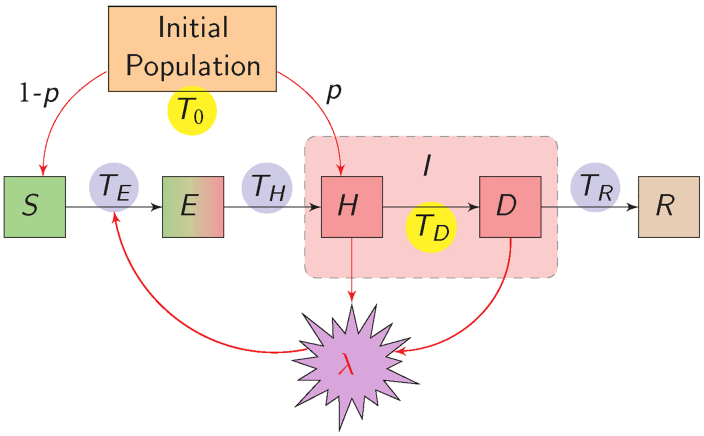

Fig 1. Susceptible-Exposed-Hidden-Detected-Removed (*SEHDR*) model of an individual's epidemiological status. At T_0, patch i is planted with infectious (I) or susceptible (S) individuals with probabilities p_i and 1-p_i, respectively. An individual passes between compartments at event times T_E, T_H, T_D and T_R. Apart from T_0, only the detection time T_D can be known (yellow); all other event times are censored (blue). Infectious individuals from both within and outside the patch contribute to the force of infection λ_{t_r}, which is the expected number of infectious events affecting an individual over time interval $(t_{r-1}, t_r]$. The probability that a given susceptible (S) individual becomes exposed (E) in this time interval is 1-exp(-λ_{t_r}), assuming independent infection events. A latent period of duration T_H-T_E follows, after which the individual becomes infectious (H). Infectious individuals are removed (R) only after detection (D) or when the entire patch is removed. For simplicity, the i and t_r subscripts are omitted in the figure.

https://doi.org/10.1371/journal.pcbi.1006085.g001

and the other for Bayesian inference–are described below (for further details, see Texts A and B in S1 Texts).

Simulation model

Whole patches are removed and replanted at predefined dates throughout the study period. Each patch i is planted with N_i individuals. At the planting date, a proportion p_i of these individuals are infectious (in state H) and 1-p_i are susceptible (in state S). If patch i is an introduction patch, $p_i>0$; otherwise, $p_i = 0$. Up to four transition times (T_E, T_H, T_D and T_R) can be associated with any given individual (Fig 1), i.e. individuals pass sequentially from state S to E to H to D to R, and all other transitions occur with zero probability. The exposed state E accounts for the latent period, i.e. the time-lag between the infection date T_E and the date at which the individual becomes infectious T_H. In this discrete-time model (whose time steps are denoted by the index r), the transitions (denoted by '→') between the five compartments are modelled as:

$$\vec{SE}_{i,t_r} \sim \text{Binom}(S_{i,t_{r-1}}, 1 - e^{-\lambda_{i,t_r}}), \tag{1}$$

$$\text{lag}(\vec{EH}) \sim \text{Gamma}_{\text{Tr}}(\theta_1, \theta_2), \tag{2}$$

$$\vec{HD}_{i,t_r} \sim \text{Binom}(H_{i,t_{r-1}}, \rho_{i,t_r}), \tag{3}$$

$$\text{lag}(\vec{DR}) \sim \text{Geom}_{\text{Tr}}(1/\delta), \tag{4}$$

where: $S_{i,t_{r-1}}$ (resp. $H_{i,t_{r-1}}$) is the number of individuals in patch i that are in state S (resp. H) at the beginning of the time interval $(t_{r-1}, t_r]$, and $\vec{SE}_{i,t_r}$ (resp. $\vec{HD}_{i,t_r}$) represents how many of them make the transition from S to E (resp. from H to D) in this time interval; the corresponding transition probabilities are $1 - e^{-\lambda_{i,t_r}}$ for a given individual in state S to incur at least one infection event (transmission of non-persistent viruses is principally driven by independent vectors), and ρ_{i,t_r} for the detection of symptoms on an infectious (H) individual ($\rho_{i,t_r} = \rho$ when patch i is inspected in $(t_{r-1}, t_r]$, and $\rho_{i,t_r} = 0$ otherwise); the sojourn times in compartments E and D are determined per individual via random variables $\text{lag}(\vec{EH}) = T_H\text{-}T_E$ and $\text{lag}(\vec{DR}) = T_R\text{-}T_D$, respectively; the latent period is modelled classically with the flexible gamma distribution, and here the left truncation of Gamma_{Tr} represents an absolute minimal latent period for sharka [33] to account for seasonality in *Prunus* phenology and prevent secondary transmission prior to the first winter; the delay between detection and removal is modelled with a geometric distribution where the probability of removal is the same ($1/\delta$) at each time step, up to the right truncation of Geom_{Tr} which represents the maximal delay before removal (detected trees must be removed before the end of the year). The force of infection (i.e. the expected number of transmission events) incurred by each individual in patch i over $(t_{r-1}, t_r]$ is defined as:

$$\lambda_{i,t_r} = \frac{\alpha_{t_r}\beta}{N_i - R_{i,t_{r-1}}} \sum_{i'} \left(m_{i'i} I_{i',t_{r-1}} \right), \tag{5}$$

where α_{t_r} is the normalized flight density, i.e. the proportion of annual flights occurring over $(t_{r-1}, t_r]$; β is the transmission coefficient, i.e. the annual number of vector flights per source (infectious) host that would lead to infection if the recipient host is susceptible; $N_i - R_{i,t_{r-1}}$ is the number of remaining hosts on which the incoming vectors distribute themselves in patch

i, and $I_{i',t_{r-1}}$ is the number of infectious hosts in patch i' over $(t_{r-1}, t_r]$. Note that N_i is constant (i.e. $N_i = S_{i,t_r} + E_{i,t_r} + I_{i,t_r} + R_{i,t_r}$) for all t_r between the planting and removal dates of patch i. Finally, the connectivity $m_{i'i}$ is the probability that a vector flight starting in patch i' terminates in patch i.

The connectivity between source patch i' of area $\mathcal{A}_{i'}$ and receiver patch i is obtained via:

$$m_{i'i} = \frac{\int_{\boldsymbol{x}\in i'}\int_{\boldsymbol{y}\in i} f^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|)d\boldsymbol{y}d\boldsymbol{x}}{\mathcal{A}_{i'}}, \tag{6}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ are coordinate vectors in $\mathbb{R}^2$, and f^{2D} is the 2-dimensional dispersal kernel [40]. The computation time required to calculate connectivity $m_{i'i}$ between several hundreds of patches prohibits the use of iterative algorithms to directly estimate the parameters of flexible (e.g. two-parameter) kernels. Thus, we developed an approach to approximate long-range (e.g. exponential-power) dispersal kernels. We defined f^{2D} as a mixture of J components:

$$f^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|) = \sum_{j=1}^{J}\left[w_j f_j^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|)\right], \tag{7}$$

where the w_j are positive mixture weights summing to 1, and $2h_j$ is the mean dispersal distance for exponential kernel f_j^{2D} defined as:

$$f_j^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|) = \frac{e^{-\|\boldsymbol{x}-\boldsymbol{y}\|/h_j}}{2\pi h_j^2}. \tag{8}$$

Under this mixture formulation, the connectivity becomes:

$$m_{i'i} = \frac{\int_{\boldsymbol{x}\in i'}\int_{\boldsymbol{y}\in i}\sum_{j=1}^{J}\left[w_j f_j^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|)\right]d\boldsymbol{y}d\boldsymbol{x}}{\mathcal{A}_{i'}} \tag{9}$$

$$= \sum_{j=1}^{J}\left[w_j \frac{\int_{\boldsymbol{x}\in i'}\int_{\boldsymbol{y}\in i} f_j^{2D}(\|\boldsymbol{x}-\boldsymbol{y}\|)d\boldsymbol{y}d\boldsymbol{x}}{\mathcal{A}_{i'}}\right]. \tag{10}$$

This formulation permits the connectivity of each mixture component j to be computed just once, since only the weights w_j require updating in an estimation procedure. We set $h_j = \frac{3}{2}\times 1.08^{j-1}$ (and $J = 100$), to obtain kernel components with mean distances ranging from 3 to 6110 m and higher resolution at smaller distances. To simplify parametrisation, and to avoid identifiability issues with the mixture of exponentials, we restrain weights using:

$$w_j = P\left(\frac{j}{J}|s_1, s_2\right) - P\left(\frac{j-1}{J}|s_1, s_2\right), \tag{11}$$

where P is the cumulative distribution function of a beta distribution with parameters s_1 and s_2. We call any kernel of the form (Eq 7) using exponential kernels (Eq 8) weighted by (Eq 11) a beta-weighted mixture of exponentials (BWME) kernel.

In order to test whether BWME kernels provide a good approximation of other dispersal kernels, we fitted a BWME kernel to 3 standard [28] dispersal kernel types (exponential-power, power-law, and 2Dt), all with the same mean distance travelled (100 m). Model fitting was performed by minimizing the total absolute difference between the marginal cumulative distribution functions at 20,000 points spaced evenly between 0 and 1000 m. For each type of disperal kernel, 4 values of the shape parameter were tested.

Bayesian estimation procedure

Among the four transition times, only T_D (i.e. the time when an infectious individual is detected) can be known precisely. Let $(t_{i,1}, \cdots, t_{i,k}, \cdots, t_{i,K_i})$ denote the set of K_i inspection dates in patch i (which may be partly censored by omissions in surveillance records). Let $p(T_{D,i} = t_{i,k})$ denote the probability for an individual in patch i to be detected as infected at inspection date $t_{i,k}$. Data provide the associated number $D^+_{i,k}$ of newly detected individuals, and the number D^-_i of individuals upon which symptoms were not detected in any of the K_i inspections. These variables are modelled as:

$$(D^+_{i,1}, \cdots, D^+_{i,K_i}, D^-_i) \sim \texttt{Multinomial}\left(N_i, p(T_{D,i} = t_{i,1}), \cdots, p(T_{D,i} = t_{i,K_i}), 1 - \sum_{k=1}^{K_i} p(T_{D,i} = t_{i,k})\right), \quad (12)$$

where N_i is the initial number of trees planted in patch i. A survival model [41] was used to derive $p(T_{D,i} = t_{i,k})$ whilst accounting for censoring, imperfect detection sensitivity, and the expected dependencies between infections (Text A in S1 Texts). The probabilities $p(T_{D,i} = t_{i,k})$ were determined from the set of model parameters Θ, using a smoothed representation of the expected epidemic, and were not conditioned on past observations. Thus, Eq (12) provides a pseudo-likelihood for the observed data (Text A in S1 Texts). Based on this pseudo-likelihood, Bayesian inference (for parameter set Θ) was performed via Markov chain Monte Carlo (MCMC) using a Gibbs sampler with embedded adaptive Metropolis-Hastings steps and data augmentation for the unknown planting and inspection dates (Texts B and C in S1 Texts). By data augmentation, we mean the explicit introduction of latent variables [42–44].

Estimation for simulated epidemics

To assess the accuracy (i.e. amount of bias) and precision (i.e. amount of variance) of the estimation of dispersal parameters, 10 epidemics were simulated under each combination of 7 disease introduction scenarios × 3 dispersal kernels × 4 parameter estimation scenarios. All simulations were performed under the same virtual landscape derived from the surveillance database: we retained the spatial coordinates (and thus the geometry) of the patch polygons, but all other potential spatio-temporal dependencies were suppressed through the random permutation of orchard-level data including planting densities and patch planting/removal/replanting dates. When density or planting date were missing in the database, their values were drawn from the corresponding empirical distribution. Simulations were performed with 1 time step per day, and 1 survey per patch per year, with inspection days drawn from the corresponding empirical distribution. The transmission coefficient β was fixed at 1.5 (which leads to realistic epidemic dynamics) and all other parameters were fixed at the expected values of their prior distributions (Text B in S1 Texts).

The three simulated kernels correspond to short-, medium- and long-range dispersal. They were parametrised using low-dimension mixtures of exponential kernels (Eq 7) with fixed mean distances and weights (Table 1, mixture parameters). These were subsequently approximated by the BWME kernel minimizing the Kullback-Leibler (KL) distance [45] between the two probability density functions (Table 1, simulation parameters).

The seven introduction scenarios were defined by the following number of introduction patches (and the initial prevalence p_i in these patches): 1 (25%), 5 (10%), 10 (5%), 15 (2%), 20 (1%), 25 (1%) or 30 (1%). For a given introduction scenario, all simulations were performed with the same introduction patches, which were chosen at random with the constraint that the first introduction occurred at year 1 and all other introductions occurred before year 6 (S1 Fig).

Table 1. Parameters of the three dispersal kernels used in the simulation study.

Kernel range	Simulation parameters		Mixture parameters	
	s_1	s_2	J	Mean distances in m (weights)
short	12727.3	29264.2	1	25 (1)
medium	9.3	18.1	2	25 (2/3), 100 (1/3)
long	5.5	8.4	3	25 (3/6), 100 (2/6), 300 (1/6)

Epidemics were simulated using BWME kernels with parameters s_1 and s_2 (left), approximating exponential mixture kernels with J mixture components (right).

https://doi.org/10.1371/journal.pcbi.1006085.t001

In order to identify whether our MCMC estimation procedure (Text C in S1 Texts) encountered identifiability issues with some parameters, we tested 4 estimation scenarios targeting parameter sets of increasing size (Table 2), with all other parameters fixed at the values used for simulation.

Both simulated epidemics and the smoothed epidemics of the pseudo-likelihood started at the beginning of year 1 and stopped at the end of year 22. Because some MCMC chains became trapped in local maxima associated with negligible likelihoods, we performed 10 MCMC chains under each estimation scenario (applied to each simulated epidemic), which produced 8400 MCMC chains in total. Within each combination of epidemic replicate × kernel × introduction × estimation scenario, we retained the MCMC chain with the highest mean posterior log-likelihood. Then, for each of these 840 chains, indices of accuracy (resp. precision) were defined as the mean (resp. span of the 95% credibility interval) of the posterior KL distances between the probability density functions f^{2D} (Eq 7) of simulated and estimated kernels. For ease of interpretation, simulated and estimated kernels were plotted using the distribution function of the distance travelled:

$$F^{1D}(\|\boldsymbol{x}-\boldsymbol{y}\|) = \sum_{j=1}^{J}\left(w_j\left[1-\left(1+\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{h_j}\right)e^{-\frac{\|\boldsymbol{x}-\boldsymbol{y}\|}{h_j}}\right]\right). \quad (13)$$

This function is the cumulative version of the 1-dimensional f^{1D} (i.e. the probability density function of the distance travelled), which is obtained by integrating (marginalising) f^{2D} (Eq 7) over all directions.

Finally, to assess the impact of detection sensitivity (ρ) on the accuracy and precision of the estimation of the dispersal kernel, we performed an additional simulation-estimation study. For 99 equally spaced values of ρ between 0.01 and 0.99, a unique epidemic was simulated. Each epidemic started at year 1 from a single introduction patch with 25% prevalence, and

Table 2. Parameter sets for four estimation scenarios.

Parameter	Definition	Θ_1	Θ_2	Θ_3	Θ_4
β	transmission coefficient	✔	✔	✔	✔
μ	$= s_1/(s_1+s_2)$; mean of kernel weight distribution	✔	✔	✔	✔
σ	$= s_1+s_2$; shape of kernel weight distribution	✔	✔	✔	✔
ρ	detection sensitivity	-	✔	-	✔
θ_1	shape of latent period distribution	-	-	✔	✔
θ_2	scale of latent period distribution	-	-	✔	✔

For each estimation scenario, the set of parameters to be estimated, Θ, comprises the parameters indicated with a ✔.

https://doi.org/10.1371/journal.pcbi.1006085.t002

spread under the long-range kernel scenario (Table 1). Default values were used for all other parameters. For each of the 99 simulated epidemics, independent estimations were carried out under the most exhaustive scheme (Θ_4) with 3 prior distributions for detection sensitivity ρ corresponding to different levels of available prior information (Text B in S1 Texts). For each combination of prior × detection sensitivity, 10 MCMC chains were run, leading to 2970 MCMC chains. For each value of ρ, posterior distributions were inferred using all chains with non-negligible mean posterior likelihood.

Estimation for a real epidemic

Using PPV surveillance data, estimation was carried out under the most exhaustive scheme (Θ_4) to infer parameters of the spatial *SEHDR* model. As above, and for the same reasons, we ran multiple MCMC chains and retained the chain with the highest mean posterior log-likelihood (Text C in S1 Texts). The number of introduction patches κ was fixed at integer values in the range 1-24, and 30 chains were run per fixed κ. This approach was taken because each unit increase in κ adds two parameters (additional introduction patch identity and initial prevalence) to Θ, which always increases the posterior log-likelihood (various uninformative and weakly informative priors were tested). Thus, to avoid over-fitting, identification of κ was treated as a model selection problem for which we maximised the Fisher information criterion $\mathcal{I}(\kappa)$ (Text D in S1 Texts).

Results

Impact of parameter values on simulated epidemics

The parameter combinations chosen to test the inference procedure cover a wide range of epidemic behaviour, from local to widespread epidemics and from low to high incidence (Fig 2). The general trends are that the stochastic variability has less effect than the introduction scenario or kernel type, that more introduction patches generally lead to more widespread epidemics, and that higher disease prevalence in the introduction patches does not necessarily increase the final local cumulative incidence (S2 and S3 Figs). Increasing kernel range generally decreases the cumulative incidence (S2 and S3 Figs), especially near the introduction patches, although these epidemics are more widespread (Fig 2).

Evaluation of the estimation procedure

A key inovation in our estimation procedure is the BWME dispersal kernel. This kernel provides close approximations to exponential-power and power-law kernels for all tested values of the shape parameter (S4 and S5 Figs). Such flexibility is an interesting property when one does not know which kernel type to assume, which is a common issue. However, the fit to the 2Dt kernels was more approximate (S6 Fig). This is not surprising since the 2Dt kernel is essentially a continuous mixture of Gaussians. Thus, switching the basis functions from exponential to Gaussian (giving a BWMG kernel) may greatly improve the fit.

The distribution of Kullback-Leibler (KL) distances between simulated and estimated kernels demonstrates that estimation accuracy is not affected by the inclusion of sensitivity and latent period parameters in the estimation scheme (Fig 3A). Neither is the median accuracy of the estimated kernels affected much by the range of the dispersal kernel (Fig 3B). However, for longer-range dispersal kernels, KL distances can become more extreme (Fig 3B), and the span and variance of their 95% credibility intervals increase (S7B Fig). This shows that the precision of the estimated kernel decreases with increasing dispersal range. The most influential factor on the accuracy and precision of estimated dispersal kernels is the introduction scenario

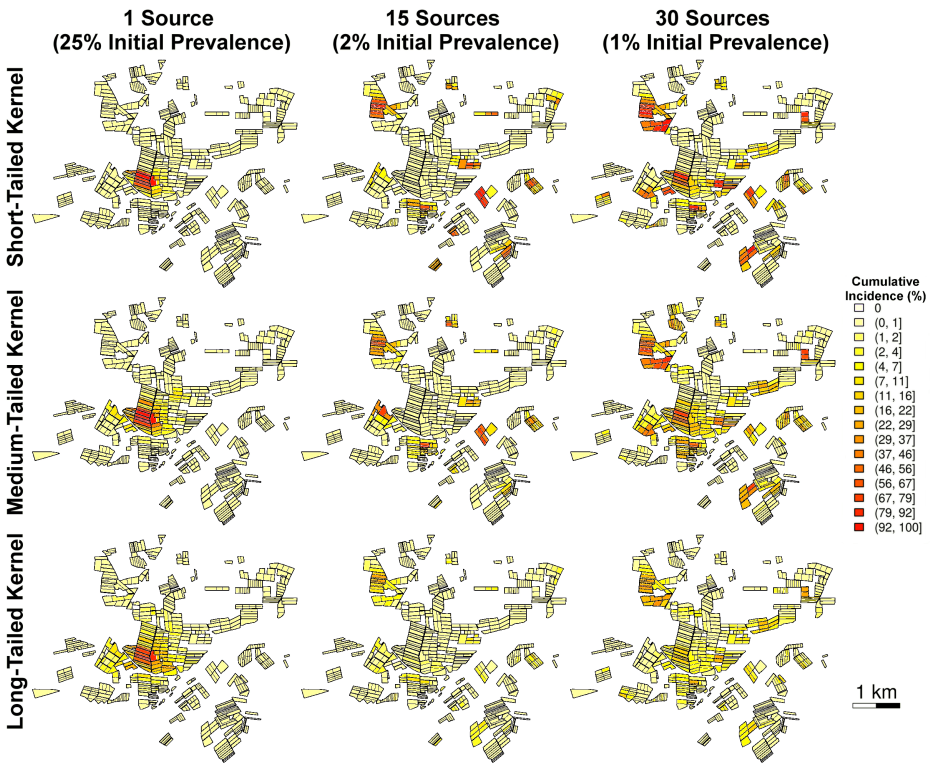

Fig 2. Cumulative detected incidence at the end of year 22 for nine simulated epidemics. Each polygon represents one peach orchard. From left to right, the number of introduction patches (with initial disease prevalence) are: 1 (25%), 15 (2%) and 30 (1%). From top to bottom: simulations generated under short-, medium- and long-range kernel scenarios.

https://doi.org/10.1371/journal.pcbi.1006085.g002

(Fig 3C and S7C Fig). However, the effect of the introduction scenario is neither strongly related to the number of introduction patches nor to the associated initial prevalence, but rather to the presence of an introduction patch in the dense central cluster of patches (Fig 3 and S1 Fig). The impact of kernel range and introduction scenario on kernel estimation can also be seen by the visual comparison between simulated and estimated kernels (S8, S9 and S10 Figs).

For each of the 3 simulated kernels, the distribution of KL distances was summarised by its minimum, quartile and maximum values across all 7 introduction scenarios × 10 epidemics per scenario. The comparison between simulated kernels and their estimates within the most exhaustive scheme (Θ_4) shows that the 3 kernels are very accurately estimated for some simulated epidemics (left column in Fig 4 and S11 Fig). However, dispersal distances are often overestimated, with the median KL distance increasing from 5.2×10^{-2} to 6.1×10^{-2} with increasing kernel range. A closer look at the estimation curves corresponding to the median KL distance reveals that estimated distances do not exceed the simulated distances by more than 0.25 on the $\log_{10}$ scale. Dispersal distances are thus overestimated by a factor below 1.8 (1.2 for the mode; see central column in S11 Fig). Even for the most challenging of the 70 epidemics simulated with the long-range dispersal kernel (bottom-right panel in Fig 4 and S11 Fig), the difference between the two curves remains below 0.6 on the $\log_{10}$ scale. This value translates into less than 4-fold estimation errors (less than 4.3 for the mode; see right column in S11 Fig), which is high but still within one order of magnitude. By contrast, precision is very high for all

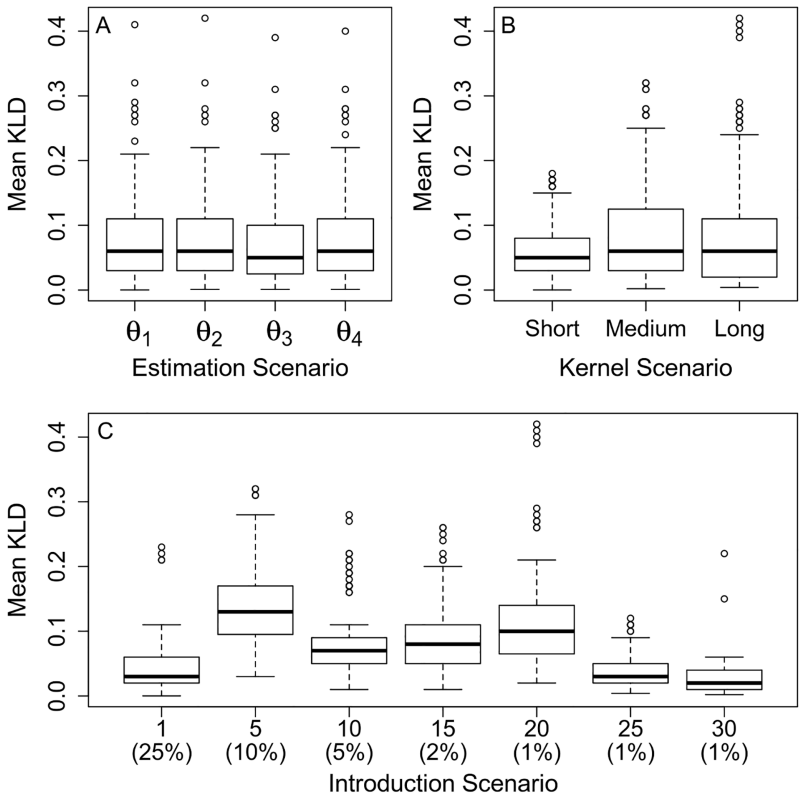

Fig 3. Boxplots of distances between simulated and estimated dispersal kernels. Impact of (A) estimation scenario, (B) kernel range, and (C) disease introduction scenario [number of introduction patches (with initial disease prevalence)] on the accuracy of estimated dispersal kernels. Accuracy is measured by the Kullback-Leibler distance (KLD) between simulated and estimated dispersal kernels. Each panel consists of 840 points, which correspond to 10 epidemics × 7 disease introduction scenarios × 3 dispersal kernels × 4 parameter estimation schemes.

https://doi.org/10.1371/journal.pcbi.1006085.g003

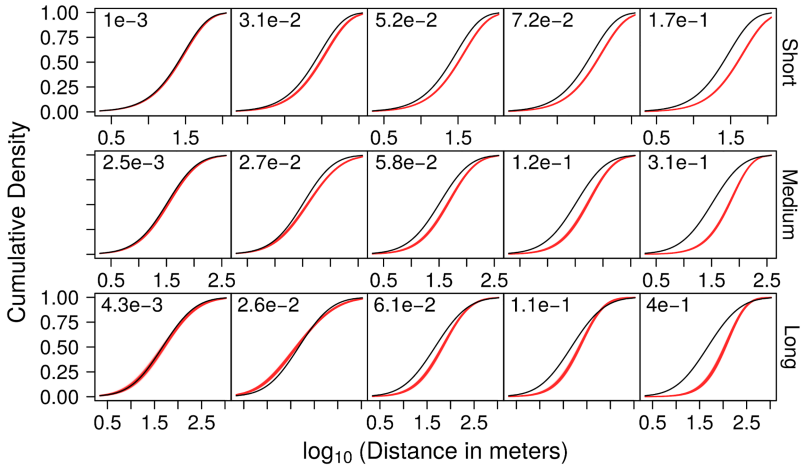

Fig 4. Comparison of simulated and estimated dispersal kernels. From left to right: kernels with the minimum, lower quartile, median, upper quartile and maximum Kullback-Leibler (KL) distances (posterior mean), as estimated (red) under the most exhaustive scheme (Θ_4), based on simulated epidemics with short-, medium- and long-range kernels (from top to bottom; black). Kernels are represented by their marginal cumulative distribution function F^{1D} (with distance from the source represented on the $\log_{10}$ scale). The mean KL distance is indicated for each estimation.

https://doi.org/10.1371/journal.pcbi.1006085.g004

kernel ranges, as indicated by a median span below 0.04 for the 95% posterior credibility interval of KL distances (S7 Fig) and the corresponding overlapping red lines in each plot of Fig 4 and S11 Fig.

The estimated values of the other parameters are generally close to the values used for simulation, but the relative bias varies among parameters, kernel ranges, and introduction scenarios (S12 Fig). Detection sensitivity (ρ) is the most precisely estimated parameter, followed by the shape of the latent period (θ_1) for which the estimates are also almost unbiased. Bias can be more severe for the scale of the latent period (θ_2) and the transmission coefficient (β), with up to 45% under- and over-estimation (respectively) in the worst-case combinations of kernel and introduction scenarios (S12 Fig, top row for θ_2 and bottom row for β). For these two parameters, the impact of the introduction scenario on parameter estimation increases with kernel range.

The simulation-estimation study on ρ shows that the estimation procedure is robust to detection sensitivities below the default value (0.8) used in the rest of this work. Indeed, although reducing ρ reduces (by definition) the proportion of detected cases, the link between detection and epidemic control results in a disproportionate increase in the total number of infected hosts as ρ decreases, providing more data for statistical inference–except when ρ reaches extremely small values (S13 Fig). As a result (see S14 and S15 Figs): (i) accuracy of kernel estimation is not reduced as detection sensitivity decreases; (ii) precision of kernel estimation is only affected when ρ is very close to 0 or 1; (iii) increasing the precision of the prior on ρ only affects the accuracy of kernel estimation for $\rho>0.8$ (i.e. when epidemic size–and thus data available for inference–is strongly reduced by effective control). Finally, we note that stochastic variations among replicated epidemics have more influence than ρ on the KL distance between simulated and estimated kernels (S15 Fig).

Estimation for a real epidemic

Once validated on simulated epidemics, we used the developed inference framework to estimate the dispersal kernel of *Plum pox virus* (and thus of the flight distances of the infectious aphid vectors) based on survey data. As a first step, we inferred the number of introduction patches. For $\kappa<10$, no combination of introduction patches returned a finite posterior log-likelihood. The Fisher information criterion was maximised at $\kappa = 11$ (Fig 5), indicating that improvement in model fit saturates beyond this point. This suggests that the most robust inference is obtained with $\kappa = 11$. These 11 introduction events among 547-579 orchards planted over 22 years (planting date is unknown for 32 orchards) correspond to disease introduction probabilities of 0.5 per year and $1.90\text{-}2.01\times10^{-2}$ per orchard planted.

Summary statistics of the posterior distributions of key parameters and percentiles of the dispersal kernel were tabulated for $\kappa = 11$ (Table 3). From the estimated values of s_1 and s_2, we derived the weights of the kernel components (S16 Fig), the dispersal kernel, the cumulative distribution function (Fig 6) and the probability density function (S17 Fig) of aphid flight distances. These figures, and the estimated quantiles shown in the second part of Table 3, demonstrate the substantial contribution of long-range dispersal to aphid-borne virus epidemics. Indeed, almost 50% of the infectious aphids leaving a tree land beyond 100 m (median distance = 92.8 m; $CI_{95\%}$ = [82.6-104 m]), and nearly 10% land beyond 1 km (last decile = 998 m; $CI_{95\%}$ = [913-1084 m]).

Discussion

In this work, we developed a spatially-explicit Bayesian inference framework for the estimation of disease dispersal parameters when surveillance data are gathered at the patch level. The

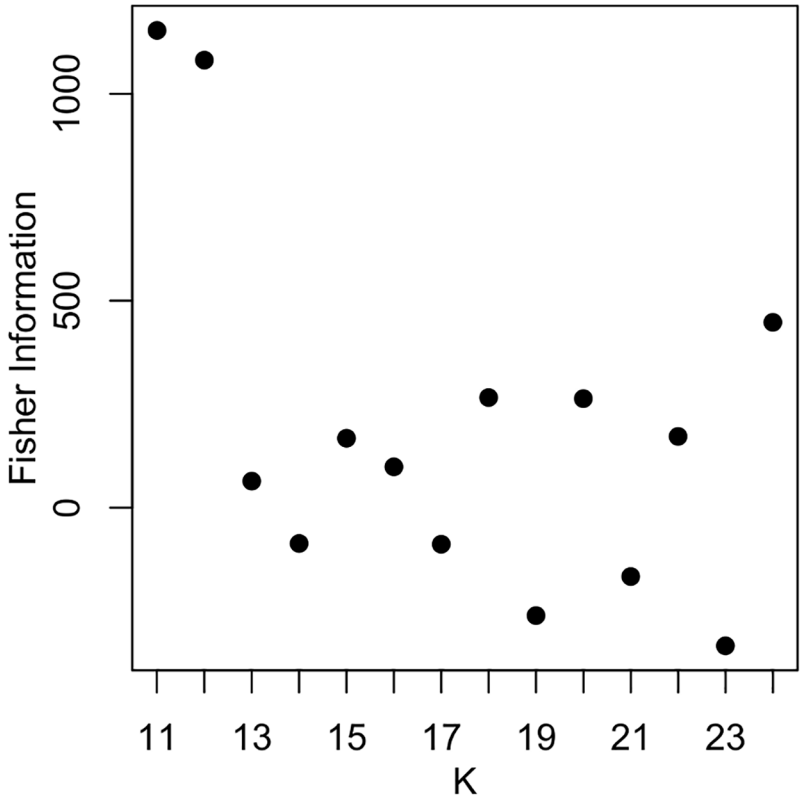

Fig 5. Impact of the number of introduction patches (κ) on the expected Fisher information for the sharka epidemic. For each κ, the estimation with the highest mean posterior log-likelihood was retained. For $\kappa<10$ no introduction patch combination returned a finite posterior log-likelihood. The empirical approximation of the Fisher information was maximal at $\kappa = 11$.

https://doi.org/10.1371/journal.pcbi.1006085.g005

Table 3. Summary statistics for parameters estimated from the survey data.

	Mean	SD	TSSEM	$CI_{95\%}$
β	1.32	2.80×10^{-2}	7.1×10^{-4}	1.27-1.38
s_1	2.32	1.11×10^{-1}	2.4×10^{-3}	2.11-2.55
s_2	2.45	8.66×10^{-2}	1.5×10^{-3}	2.29-2.62
ρ	0.659	7.73×10^{-3}	1.9×10^{-4}	0.643-0.674
θ_{exp}	1.92	8.74×10^{-2}	2.2×10^{-3}	1.75-2.09
θ_{var}	0.442	8.69×10^{-2}	2.0×10^{-3}	0.291-0.631
$d_{5\%}$	5.0	3.23×10^{-1}	7.2×10^{-3}	4.4-5.7
$d_{10\%}$	8.9	5.98×10^{-1}	1.3×10^{-2}	7.8-10.1
$d_{50\%}$	92.8	5.47	1.2×10^{-1}	82.6-104
$d_{90\%}$	998	4.35×10^{1}	7.7×10^{-1}	913-1084
$d_{95\%}$	1742	7.20×10^{1}	1.2	1604-1887

Summary statistics including the posterior mean, standard deviation (SD), time-series standard error of the mean (TSSEM) and 95% credibility intervals ($CI_{95\%}$) are reported for the transmission coefficient β, the kernel parameters s_1 and s_2, the detection sensitivity ρ, the expected duration $\theta_{exp} = \theta_1 \times \theta_2$ of the latent period and associated variance $\theta_{var} = \theta_1 \times \theta_2^2$. Posterior distributions of the 5th, 10th, 50th, 90th and 95th percentiles of aphid flight distances d are also summarised.

https://doi.org/10.1371/journal.pcbi.1006085.t003

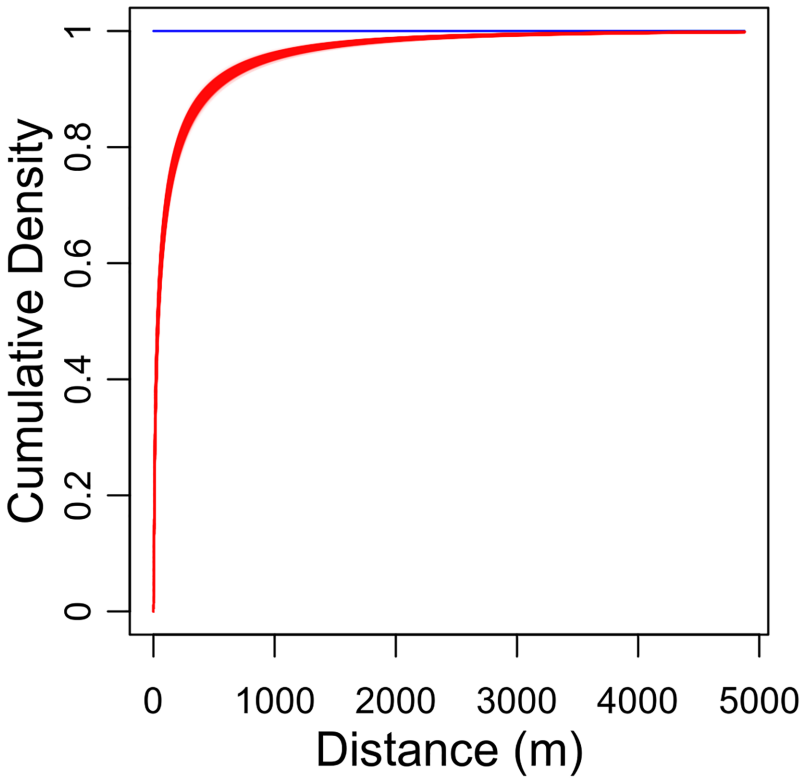

Fig 6. Estimated dispersal kernel for the sharka epidemic. The posterior marginal cumulative distribution function, F^{1D}, of the fitted dispersal kernel, obtained for $\kappa = 11$ (i.e. the number of introduction patches maximising the Fisher information). The plotted posterior distribution was obtained from 4000 MCMC samples. One line is plotted per sample.

https://doi.org/10.1371/journal.pcbi.1006085.g006

simulation and inference procedures take into account that disease status assessment is incomplete because surveillance has an imperfect detection sensitivity and a finite spatio-temporal coverage. We assessed the quality of the inference procedure through comparison between parameter values used for simulation and corresponding estimates. Then, we applied this approach to *Plum pox virus* surveillance data, to obtain the first estimate of an aphid dispersal kernel at the landscape scale. We discuss below the interest and limitations of the proposed approach and results.

Sources of uncertainty and model validation

Since the dispersal kernel is the key component of spatial epidemiological models, we focused attention on its estimation and treated the other parameters as nuisance parameters (i.e. parameters than are inferred to limit bias in the estimation of the distribution of interest). S12 Fig shows how simulated and estimated values compare for all nuisance parameters. Recent methodological advances have permitted the extraction of crucial information on the dispersal kernel of four plant diseases from surveillance data [13, 16–18] and observational studies [15, 30]. These estimation procedures all account for unobserved infection times, with additional methodological challenges related to large heterogeneous landscapes [16], introduction from external sources [17, 18], or active disease control [18]. The present work handles these various processes and, contrary to the abovementioned studies which all assume a known detection

sensitivity, also accounts for this poorly known variable which adds a layer of uncertainty into the surveillance process. Inclusion of parameters for detection sensitivity and the latent period in the estimation procedure (Table 2) barely affects the KL distance between simulated and estimated kernels (Fig 3 and S7 Fig); hence the inclusion of these extra parameters during inference based on PPV surveillance data. The resulting estimate of detection sensitivity is ρ = 0.66 (Table 3). Although a previous analysis showed that the presence of undetected infectious individuals resulted in slightly overestimated dispersal distances [19], here we show that our estimation procedure is robust to detection sensitivities far below one, even with weak prior information on ρ (S14 and S15 Figs). The dataset used for inference contains information on the disease status of more than 401,000 trees over 15 years, and is associated with a substantial level of censoring (on the dates of planting, inspection, infection, end of the latent period, and removal). For these reasons, using data augmentation to infer the transition times was unlikely to scale successfully to our analysis. Instead, we used a pseudo-likelihood where the unknown numbers of infectious and removed trees were replaced by their expected values. Intuitively, this approach can be expected to work best in highly connected landscapes, where epidemics are more likely to follow their expected course, and to become more erroneous in patchy landscapes where stochastic events can deflect epidemics away from their expected course. This might explain in part why the smaller KL distances in Fig 3C correspond to those introduction scenarios where a source patch was located in the most highly connected region of the study area.

A unique feature of the present work is the validation of the estimation of the dispersal kernel through comparing known functions used in simulations and the corresponding functions estimated from these simulated epidemiological data sets. Although this is an intuitive and standard practice [46–48], previous estimations of plant disease dispersal parameters instead used goodness-of-fit statistics between actual and simulated spatiotemporal patterns as a way to validate their inference models [16–18]. This general trend to rely on goodness-of-fit statistics, without performing simulation-based validation tests, may be due to the high computational burden associated with such validation procedures which require several simulation scenarios and several independent estimations per scenario to assess the accuracy and precision of the estimation algorithms. Since we focus on dispersal kernel estimation, rather than on model predictions as in [16, 17], simulation-based validation was useful to demonstrate that, despite the approximations of the pseudo-likelihood, dispersal kernel estimation was generally very precise. Accuracy was often high for short-range kernels, and dispersal distance estimates ranged from very accurate to overestimated for longer-range kernels (Figs 3 and 4). The same approach also showed that both the precision and the accuracy of dispersal kernel estimation is unaltered when the probability ρ to detect a symptomatic/infectious tree is in the range 0.05–0.8 (S15 Fig).

The observed overestimation is not likely to be caused by insufficient flexibility in the BWME kernel because, even for the 2Dt dispersal kernel (which the BWME kernel does not fit perfectly), the magnitude of the difference between the two kernels is negligible in comparison with the difference between simulated and estimated kernels. It is not likely either to be caused by choosing the MCMC chain with the highest mean posterior likelihood (among 10 chains) since this procedure was just used to remove degenerate chains (and coherence between all other chains was high). Although this procedure is rather wasteful of problem-free chains, and provides lower precision than alternative approaches to multi-chain analysis, there is no reason to expect any bias concerning the mean (or other statistics) of the posterior distribution. It is most likely that the estimation bias reported here arose from approximations made (for practical reasons) within the pseudo-likelihood.

Dispersal in a patchy landscape

Our inference procedure explicitly accounts for patch geometry and patch-level aggregation of surveillance data. Although this choice was data-driven (infected tree numbers–not individual locations–were included in the database), for landscape-scale studies this approach appears to strike an interesting compromise between computational feasibility and spatial realism. Indeed, considering the disease status of over 401,000 individuals simultaneously would cause major computational issues given the size of the resulting connectivity matrix. Conversely, spatial models commonly use the coordinates of patch centroids in connectivity calculations (e.g. [14, 19]). However, this neglects patch geometry and can be expected to bias connectivity estimates (i) when patch shapes and sizes are disparate, or (ii) when patch dimensions are of the same order of magnitude as the distances between patches. To exemplify (i), consider a small patch located next to a large patch, where many of the propagules leaving the small patch can be expected to land, but a much lower proportion of the propagules leaving the large patch are expected to fall in the small patch. To exemplify the importance of (ii), consider that many more propagules can be exchanged between two large adjacent orchards than would be calculated using the distance between their distant centroids. Although our approach neglects the effects of disease aggregation within patches, it does account for patch size and geometry that both impact disease spread [49]. The use of Eq (6) to integrate patch geometry, combined with the BWME kernel, can thus be useful for the inference of the landscape-scale dispersal kernels of many wind- and vector-borne diseases.

A rigourous assessment of connectivity between patches is also necessary because of its influence on parameter estimation. Our study shows that kernel range affects both the KL distance between simulated and estimated dispersal kernels (Fig 3 and S7 Fig) and the cumulative incidence (S2 and S3 Figs). This pattern reflects how parameter identifiability depends on statistical power, which depends on cumulative disease incidence, which in turn depends on landscape connectivity. Short-range kernels imply greater local connectivity than long-range kernels, leading to relatively intense local transmission but reduced transmission at greater distances. Whether or not shorter-range kernels generate larger epidemics depends on the proportion of potential transmission events falling outside host patches, and thus on landscape configuration. Here, larger cumulative incidences were obtained using smaller kernels because, in our patchy agricultural landscape, many dispersal events generated by long-tailed kernels do not end within host patches.

Impact of disease introductions on inference

Disease introduction scenarios had a substantial effect on the accuracy and precision of the inferred dispersal kernel (Fig 3 and S7 Fig). Surprisingly, this effect does not seem related to either the number of introduction patches or the associated initial prevalence. However, we note that lower KL distances between simulated and estimated dispersal kernels (in introduction scenarios 1, 6 and 7) are associated with introductions occurring in the highly connected central patches (S1 Fig). The resulting higher cumulative incidence probably improves estimation for the reasons given above.

During parameter estimation, we did encounter multi-modality in the posterior likelihood surface, which may arise when fitting ecological dynamic models to data, even without observation error and model mis-specification [50]. For epidemic scenarios with both a short-range kernel and a high number of introduction events, misidentifying some of the introduction patches had a large negative effect on the likelihood, and some MCMC chains were trapped in degenerate solutions. For this reason, we ran the MCMC algorithms many times and carefully compared the posterior likelihoods and parameter estimates of all chains before making inference. We also

considered alternative algorithms such as parallel tempering [51] or equi-energy sampling [52], which increase the likelihood of between-mode transitions. However, the extra computational burden of these approaches was considered superfluous given that the observed differences in the posterior likelihoods of various modes were typically relatively large. Thus, launching a large number of chains to increase the likelihood of identifying the global mode was a reasonable compromise. We have extensively tested this approach, reporting here the results of several thousand MCMC chains, and have found that in practice results are consistent.

Overall, inference of epidemiological parameters is easier for epidemics where disease introductions are well characterized, or at least infrequent. Unfortunately, this was not the case with the PPV-M dataset, and estimating the number of introduction patches κ was challenging. Such difficulty is by no means unique to the current study (see e.g. [17]). Reversible-jump MCMC (RJMCMC) [53] is a method for performing MCMC when the dimension of the parameter space is unknown and inferred from data. We initially attempted various implementations of RJMCMC, but found it impossible to construct priors that could both prevent over-fitting and provide robust posterior probabilities for κ under a wide variety of epidemiological scenarios. To circumvent this issue we inferred κ based on the Fisher information. This gives a minimum-variance estimator that provides robust inference with a good balance between under- and over-fitting–although it does not permit the estimation of posterior probabilities associated with the various κ. This approach has been used successfully in similar situations [54].

Insights into aphid biology

Like most plant viruses, PPV is transmitted by winged non-colonising aphids in a non-persistent manner [33]. To match the characteristics of this widespread transmission process, in our model transmission events are independent (conditional on infection sources) and transmission distances directly depend on host locations and on the distance travelled by an aphid within a single infectious flight. Although estimating this aphid dispersal kernel is crucial to plant virus epidemiology, it has long remained elusive. Traditional ecological methods such as capture-mark-recapture provide little information regarding aphid dispersal at the landscape scale [32]. This has been a major obstacle to the parametrisation of models simulating the dispersal of these vectors and the pathogens they spread, as exemplified by the scarcity of landscape-scale models on cereal aphids [55] and by the informed guesses of flight-distance parameters in such models [56]. Here we estimated, for the first time, the dispersal of aphid vectors at the landscape scale. This estimation indicates that 50% of the infectious aphids leaving a tree land within about 90 meters, while about 10% of flights terminate beyond 1 km. Although dispersal estimation from simulated epidemics suggests that these distances may be overestimated, the large number of flights estimated to terminate within some tens of meters of the source tree is consistent with previous studies of within-patch clustering of trees infected by PPV-M [33, 57, 58] or PPV-D [38, 59]. Indeed, one of these studies [38] shows that 50% of the new PPV cases occur within 35-70 m of the nearest previous case; in addition, 10% of the new PPV cases were found beyond 200-460 m from the nearest previous case. Although the proportion of new PPV cases captured within a given radius is not equivalent to a dispersal kernel (e.g. because the trees are not always infected by the nearest previously detected neighbour), the figures are of the same order of magnitude. In particular, both studies highlight the long range of the dispersal kernel. Our estimation of the dispersal kernel at the landscape scale has important consequences. For example, current French regulations enforce at least one visual inspection per year within 2.5 km of a detected sharka case (followed by the removal of all trees with sharka symptoms). Our results suggest that less than 3% of flights should thus go

beyond this radius (Fig 6). In a patchy French landscape, most of these aphids would land outside a peach orchard and thus lead to no infection. Such procedures are thus likely to efficiently detect most of the aphid-mediated secondary infections; actually, given the cost of surveillance and the speed of disease spread, this radius may even be oversized. Future work based on this study could aim at the definition of new management strategies against PPV. More generally, our results provide a unique reference point on the epidemiology, simulation and control of the principal group of plant viruses (i.e. those caused by non-persistant aphid-borne viruses), which have a major epidemiological and economic impact. Finally, by focusing on incidence data the presented estimation approach is adaptable to many epidemiological situations, including other vector-borne and airborne fungal diseases.

Supporting information

S1 Fig. Simulated planting years and introduction patch locations used in the simulation study. The first map (top left) represents the randomisation of the first planting years of the 553 patches. These years were sampled without replacement from their empirical distribution. The other maps show the location and planting year of each introduction patch in the seven introduction scenarios. The number of introduction patches and their initial prevalence are indicated for each introduction scenario. Note the greater landscape connectivity in the central area.
(TIFF)

S2 Fig. Cumulative detected incidence for the introduction scenarios tested in the simulation study. For each introduction scenario, the number of introduction patches and the corresponding initial disease prevalence are mentionned above the graph. The three tested kernels are represented by different colours. For each combination of kernel and introduction scenarios, 10 independant simulated epidemics are shown.
(TIFF)

S3 Fig. Cumulative detected incidence for the kernels tested in the simulation study. For each kernel, the seven tested introduction scenarios are represented by different colours. For each combination of kernel and introduction scenarios, 10 independant simulated epidemics are shown.
(TIFF)

S4 Fig. Best-fit BWME kernel approximations of exponential-power kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.
(TIFF)

S5 Fig. Best-fit BWME kernel approximations of power-law kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.
(TIFF)

S6 Fig. Best-fit BWME kernel approximations of 2Dt kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.
(TIFF)

S7 Fig. Boxplots of the variation among estimated dispersal kernels. Impact of (A) estimation scenario, (B) kernel range, and (C) disease introduction scenario [number of introduction patches (with initial disease prevalence)] on the precision of estimated dispersal kernels. Precision is measured by the span of the 95% credibility interval of Kullback-Leibler distances (Span KLD) between simulated and estimated dispersal kernels. Each panel consists of 840 points, which correspond to 10 epidemics × 7 disease introduction scenarios × 3 dispersal kernels × 4 parameter estimation schemes.
(TIFF)

S8 Fig. Influence of introduction scenarios on the estimation of a short-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a short-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood (mean proportion: 10%) increases quadratically with the number of source orchards. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).
(TIFF)

S9 Fig. Influence of introduction scenarios on the estimation of a medium-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a medium-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood varies among introduction scenarios, with a mean proportion of 2.6%. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).
(TIFF)

S10 Fig. Influence of introduction scenarios on the estimation of a long-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a long-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood is low (mean proportion: 0.4%) for all the introduction scenarios. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).
(TIFF)

S11 Fig. Comparison of simulated and estimated dispersal kernels. From left to right: kernels with the minimum, lower quartile, median, upper quartile and maximum Kullback-Leibler (KL) distances (posterior mean), for all chains with non-negligible mean posterior likelihood. Estimations (red) under the most exhaustive scheme (Θ_4) are based on simulated epidemics with short-, medium- and long-range kernels (from top to bottom; black). Kernels

are represented by their marginal probability density function f^{1D}. The mean KL distance is indicated for each estimation.
(TIFF)

S12 Fig. Comparison of simulated and estimated nuisance parameters. For each combination of short-, medium- and long-range kernels (from top to bottom) and introduction scenarios (colour-coded as in S3, S8, S9 and S10 Figs), 10 epidemics were simulated and 10 MCMC chains were run per simulated epidemic. The curves represent the posterior distribution of the parameters obtained under the most exhaustive estimation scheme (Θ_4) for all chains with non-negligible mean posterior likelihood. Dashed lines: parameter values used in the simulations.
(TIFF)

S13 Fig. Cumulative detected incidence at the end of year 22 across the range of detection sensitivities (ρ) tested in the dedicated simulation study. Each polygon represents one peach orchard. All eight simulations start at year 1 from a unique introduction patch with 25% initial prevalence and spread is determined by the long-range kernel. Note that the final detected prevalence varies non-monotonically with detection sensitivity because the removal of detected trees reduces disease spread.
(TIFF)

S14 Fig. Influence of detection sensitivity on the estimation of the long-range dispersal kernel. For each detection sensitivity, a single epidemic was simulated using the long-range kernel (black dashed curve). The posterior distributions of the estimated kernels (obtained from all MCMC chains with non-negligible mean posterior likelihood) are shown for three levels of prior information. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).
(TIFF)

S15 Fig. Influence of detection sensitivity on the distance between simulated and estimated long-range dispersal kernels. For each of the 99 detection sensitivities, a single epidemic was simulated using the long-range kernel. For three levels of prior information, each bar represents a 95% credibility interval on the Kullback-Leibler distance (KLD) between simulated and estimated dispersal kernels (obtained from all MCMC chains with non-negligible mean posterior likelihood). The grey vertical lines correspond to the values of detection sensitivity used in S13 and S14 Figs.
(TIFF)

S16 Fig. Estimated weights of the (BWME) dispersal kernel for the sharka epidemic. The posterior distribution of the weights (calculated with (Eq 11) for a mixture of 100 exponential kernels) is obtained for $\kappa = 11$ (i.e. the number of introduction patches maximising the Fisher information). The plotted posterior distribution of weights (as a function of the expected distance of each kernel) was obtained from 4000 MCMC samples. One line is plotted per sample.
(TIFF)

S17 Fig. Estimated dispersal density for the sharka epidemic. The posterior distribution of the marginal probability density function, f^{1D}, of the fitted dispersal kernel, obtained for $\kappa = 11$ (i.e. the number of introduction patches maximising the Fisher information). The plotted posterior distributions were obtained from 4000 MCMC samples. One line is plotted per sample.
(TIFF)

S1 Texts. (A) Probabilistic framework for statistical inference, (B) prior distributions, (C) Markov chain Monte Carlo, and (D) model selection for κ.
(PDF)

Acknowledgments

The PPV-M dataset was provided by the "Fédération départementale de défense contre les organismes nuisibles". Sylain Grizard provided invaluable technical assistance with geographical information systems. We are also grateful to Annie Bouvier (INRA) for her technical support regarding the software CaliFloPP, and to Eric Montaudon and Véronique Martin (INRA) for their help with the MIGALE computer cluster.

Author Contributions

Conceptualization: David R. J. Pleydell, Samuel Soubeyrand, Gérard Labonne, Joël Chadœuf, Gaël Thébaud.

Data curation: David R. J. Pleydell, Sylvie Dallot, Gérard Labonne, Gaël Thébaud.

Formal analysis: David R. J. Pleydell.

Funding acquisition: David R. J. Pleydell, Sylvie Dallot, Gérard Labonne, Emmanuel Jacquot, Gaël Thébaud.

Investigation: David R. J. Pleydell, Sylvie Dallot, Gérard Labonne, Gaël Thébaud.

Methodology: David R. J. Pleydell, Samuel Soubeyrand, Joël Chadœuf, Gaël Thébaud.

Project administration: Gérard Labonne, Emmanuel Jacquot, Gaël Thébaud.

Resources: Sylvie Dallot, Gérard Labonne.

Software: David R. J. Pleydell.

Supervision: Gérard Labonne, Emmanuel Jacquot, Gaël Thébaud.

Validation: David R. J. Pleydell, Gaël Thébaud.

Visualization: David R. J. Pleydell, Gaël Thébaud.

Writing – original draft: David R. J. Pleydell, Gaël Thébaud.

Writing – review & editing: David R. J. Pleydell, Samuel Soubeyrand, Sylvie Dallot, Gérard Labonne, Joël Chadœuf, Emmanuel Jacquot, Gaël Thébaud.

References

1. Anderson RM, May RM. Infectious Diseases of Humans Dynamics and Control. Oxford University Press; 1992. Available from: http://www.oup.com/uk/catalogue/?ci=9780198540403.
2. Keeling MJ, Rohani P. Modeling Infectious Diseases in Humans and Animals. Princeton University Press; 2007.
3. Savary S. Epidemics of Plant Diseases: Mechanisms, Dynamics and Management. In: Tibayrenc M, editor. Encyclopedia of Infectious Diseases. John Wiley & Sons, Inc.; 2007. p. 125–136.
4. Parnell S, Gottwald TR, Gilligan CA, Cunniffe NJ, van den Bosch F. The effect of landscape pattern on the optimal eradication zone of an invading epidemic. Phytopathology. 2010; 100(7):638–644. https://doi.org/10.1094/PHYTO-100-7-0638 PMID: 20528181
5. Keeling MJ, Brooks SP, Gilligan CA. Using conservation of pattern to estimate spatial parameters from a single snapshot. Proc Natl Acad Sci USA. 2004; 101(24):9155–9160. https://doi.org/10.1073/pnas.0400335101 PMID: 15184669

6. Rorres C, Pelletier STK, Keeling MJ, Smith G. Estimating the kernel parameters of premises-based stochastic models of farmed animal infectious disease epidemics using limited, incomplete, or ongoing data. Theor Popul Biol. 2010; 78(1):46–53. https://doi.org/10.1016/j.tpb.2010.04.003 PMID: 20452368

7. Cunniffe NJ, Koskella B, Metcalf CJE, Parnell S, Gottwald TR, Gilligan CA. Thirteen challenges in modelling plant diseases. Epidemics. 2015; 10:6–10. https://doi.org/10.1016/j.epidem.2014.06.002 PMID: 25843374

8. Gerbier G, Bacro J, Pouillot R, Durand B, Moutou F, Chadœuf J. A point pattern model of the spread of foot-and-mouth disease. Prev Vet Med. 2002; 56(1):33–49. https://doi.org/10.1016/S0167-5877(02)00122-8 PMID: 12419598

9. Gibson GJ, Kleczkowski A, Gilligan CA. Bayesian analysis of botanical epidemics using stochastic compartmental models. Proc Natl Acad Sci USA. 2004; 101(33):12120–12124. https://doi.org/10.1073/pnas.0400829101 PMID: 15302941

10. Cook AR, Otten W, Marion G, Gibson GJ, Gilligan CA. Estimation of multiple transmission rates for epidemics in heterogeneous populations. Proc Natl Acad Sci USA. 2007; 104(51):20392–20397. https://doi.org/10.1073/pnas.0706461104 PMID: 18077378

11. Neri FM, Bates A, Füchtbauer WS, Pérez-Reche FJ, Taraskin SN, Otten W, et al. The effect of heterogeneity on invasion in spatial epidemics: From theory to experimental evidence in a model system. PLoS Comput Biol. 2011; 7(9):e1002174. https://doi.org/10.1371/journal.pcbi.1002174 PMID: 21980273

12. Ludlam JJ, Gibson GJ, Otten W, Gilligan CA. Applications of percolation theory to fungal spread with synergy. J R Soc Interface. 2012; 9(70):949–956. https://doi.org/10.1098/rsif.2011.0506 PMID: 22048947

13. Cunniffe NJ, Laranjeira FF, Neri FM, DeSimone RE, Gilligan CA. Cost-effective control of plant disease when epidemiological knowledge is incomplete: modelling Bahia bark scaling of citrus. PLoS Comput Biol. 2014; 10(8):e1003753. https://doi.org/10.1371/journal.pcbi.1003753 PMID: 25102099

14. Chis Ster I, Singh BK, Ferguson NM. Epidemiological inference for partially observed epidemics: the example of the 2001 foot and mouth epidemic in Great Britain. Epidemics. 2009; 1(1):21–34. https://doi.org/10.1016/j.epidem.2008.09.001 PMID: 21352749

15. Soubeyrand S, Laine AL, Hanski I, Penttinen A. Spatiotemporal structure of host-pathogen interactions in a metapopulation. Am Nat. 2009; 174(3):308–320. https://doi.org/10.1086/603624 PMID: 19627233

16. Meentemeyer RK, Cunniffe NJ, Cook AR, Filipe JA, Hunter RD, Rizzo DM, et al. Epidemiological modeling of invasion in heterogeneous landscapes: spread of sudden oak death in California (1990–2030). Ecosphere. 2011; 2(2):1–24. https://doi.org/10.1890/ES10-00192.1

17. Neri FM, Cook AR, Gibson GJ, Gottwald TR, Gilligan CA. Bayesian analysis for inference of an emerging epidemic: citrus canker in urban landscapes. PLoS Comput Biol. 2014; 10(4):e1003587. https://doi.org/10.1371/journal.pcbi.1003587 PMID: 24762851

18. Parry M, Gibson GJ, Parnell S, Gottwald TR, Irey MS, Gast TC, et al. Bayesian inference for an emerging arboreal epidemic in the presence of control. Proc Natl Acad Sci USA. 2014; 111(17):6258–6262. https://doi.org/10.1073/pnas.1310997111 PMID: 24711393

19. Salje H, Lessler J, Paul KK, Azman AS, Rahman MW, Rahman M, et al. How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. Proc Natl Acad Sci USA. 2016; p. 13420–13425. https://doi.org/10.1073/pnas.1611391113 PMID: 27821727

20. Soubeyrand S, Thébaud G, Chadœuf J. Accounting for biological variability and sampling scale: a multi-scale approach to building epidemic models. J Roy Soc Interface. 2007; 4(16):985–997. https://doi.org/10.1098/rsif.2007.1154

21. King AA, Ionides EL, Pascual M, Bouma MJ. Inapparent infections and cholera dynamics. Nature. 2008; 454(7206):877–880. https://doi.org/10.1038/nature07084 PMID: 18704085

22. Birrell PJ, Ketsetzis G, Gay NJ, Cooper BS, Presanis AM, Harris RJ, et al. Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London. Proc Natl Acad Sci USA. 2011; 108(45):18238–18243. https://doi.org/10.1073/pnas.1103002108 PMID: 22042838

23. Gambino B. The correction for bias in prevalence estimation with screening tests. J Gambl Stud. 1997; 13(4):343–351. https://doi.org/10.1023/A:1024971521887 PMID: 12913383

24. Thompson RN, Cobb RC, Gilligan CA, Cunniffe NJ. Management of invading pathogens should be informed by epidemiology rather than administrative boundaries. Ecol Model. 2016; 324:28–32. https://doi.org/10.1016/j.ecolmodel.2015.12.014

25. Box-Steffensmeier JM, Jones BS. Event History Modeling: A Guide for Social Scientists. Cambridge University Press; 2004.

26. Zhang XS, Holt J, Colvin J. A general model of plant-virus disease infection incorporating vector aggregation. Plant Pathol. 2000; 49(4):435–444. https://doi.org/10.1046/j.1365-3059.2000.00469.x

27. Austerlitz F, Dick CW, Dutech C, Klein EK, Oddou-Muratorio S, Smouse PE, et al. Using genetic markers to estimate the pollen dispersal curve. Mol Ecol. 2004; 13(4):937–954. https://doi.org/10.1111/j.1365-294X.2004.02100.x PMID: 15012767

28. Nathan R, Klein E, Robledo-Arnuncio JJ, Revilla E. Dispersal kernels: review. In: Clobert J, Baguette M, Benton TG, Bullock JM, editors. Dispersal Ecology and Evolution. Oxford University Press; 2012. p. 187–210.

29. Rieux A, Soubeyrand S, Bonnot F, Klein EK, Ngando JE, Mehl A, et al. Long-distance wind-dispersal of spores in a fungal plant pathogen: estimation of anisotropic dispersal kernels from an extensive field experiment. PLoS ONE. 2014; 9(8):e103225. https://doi.org/10.1371/journal.pone.0103225 PMID: 25116080

30. Bousset L, Jumel S, Garreta V, Picault H, Soubeyrand S. Transmission of *Leptosphaeria maculans* from a cropping season to the following one. Ann Appl Biol. 2015; 166(3):530–543. https://doi.org/10.1111/aab.12205

31. Nault L. Arthropod transmission of plant viruses: a new synthesis. Ann Entomol Soc America. 1997; 90(5):521–541. https://doi.org/10.1093/aesa/90.5.521

32. Loxdale HD, Hardie J, Halbert S, Footit R, Kidd NA, Carter CI. The relative importance of short- and long-range movement of flying aphids. Biol Rev. 1993; 68(2):291–311. https://doi.org/10.1111/j.1469-185X.1993.tb00998.x

33. Rimbaud L, Dallot S, Gottwald T, Decroocq V, Jacquot E, Soubeyrand S, et al. Sharka epidemiology and worldwide management strategies: Learning lessons to optimize disease control in perennial plants. Annu Rev Phytopathol. 2015; 53:357–378. https://doi.org/10.1146/annurev-phyto-080614-120140 PMID: 26047559

34. Scholthof KBG, Adkins S, Czosnek H, Palukaitis P, Jacquot E, Hohn T, et al. Top 10 plant viruses in molecular plant pathology. Molec Plant Pathol. 2011; 12(9):938–954. https://doi.org/10.1111/j.1364-3703.2011.00752.x

35. Cambra M, Capote N, Myrta A, Llácer G. *Plum pox virus* and the estimated costs associated with sharka disease. EPPO Bull. 2006; 36(2):202–204. https://doi.org/10.1111/j.1365-2338.2006.01027.x

36. Labonne G, Yvon M, Quiot JB, Avinent L, Llácer G. Aphids as potential vectors of plum pox virus: comparison of methods of testing and epidemiological consequences. Acta Hortic. 1995; 386:207–218. https://doi.org/10.17660/ActaHortic.1995.386.27

37. Perring TM, Gruenhagen NM, Farrar CA. Management of plant viral diseases through chemical control of insect vectors. Annu Rev Entomol. 1999; 44(1):457–481. https://doi.org/10.1146/annurev.ento.44.1.457 PMID: 15012379

38. Gottwald TR, Wierenga E, Luo W, Parnell S. Epidemiology of Plum pox 'D' strain in Canada and the USA. Canad J Plant Pathol. 2013; 35(4):442–457. https://doi.org/10.1080/07060661.2013.844733

39. Quiot JB, Labonne G, Boeglin M, Adamolle C, Renaud LY, Candresse T. Behaviour of two isolates of plum pox virus inoculated on peach and apricot trees: first results. Acta Hortic. 1995; 386:290–297. https://doi.org/10.17660/ActaHortic.1995.386.39

40. Bouvier A, Kiêu K, Adamczyk K, Monod H. Computation of the integrated flow of particles between polygons. Environ Model Softw. 2009; 24(7):843–849. https://doi.org/10.1016/j.envsoft.2008.11.006

41. Cox DR, Oakes D. Analysis of survival data. Chapman & Hall; 1984.

42. Tanner MA, Wong WH. The calculation of posterior distributions by data augmentation. JASA. 1987; 82(398):528–540. https://doi.org/10.1080/01621459.1987.10478458

43. van Dyk DA, Meng XL. The art of data augmentation. J Comput Graph Stat. 2001; 10(1):1–50. https://doi.org/10.1198/10618600152418584

44. McKinley TJ, Ross JV, Deardon R, Cook AR. Simulation-based Bayesian inference for epidemic models. Comput Stat Data Anal. 2014; 71:434–447. https://doi.org/10.1016/j.csda.2012.12.012

45. Kullback S, Leibler RA. On Information and Sufficiency. Ann Math Statist. 1951; 22(1):79–86. https://doi.org/10.1214/aoms/1177729694

46. Lehmann EL, Casella G. Theory of point estimation. Springer Science & Business Media; 2006.

47. Morelli MJ, Thébaud G, Chadœuf J, King DP, Haydon DT, Soubeyrand S. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. PLoS Comput Biol. 2012; 8(11):e1002768. https://doi.org/10.1371/journal.pcbi.1002768 PMID: 23166481

48. Lau MS, Marion G, Streftaris G, Gibson G. A systematic Bayesian integration of epidemiological and genetic data. PLoS Comput Biol. 2015; 11(11):e1004633. https://doi.org/10.1371/journal.pcbi.1004633 PMID: 26599399

49. Mikaberidze A, Mundt CC, Bonhoeffer S. Invasiveness of plant pathogens depends on the spatial scale of host distribution. Ecol Appl. 2016; 26(4):1238–1248. https://doi.org/10.1890/15-0807 PMID: 27509761

50. Polansky L, de Valpine P, Lloyd-Smith JO, Getz WM. Likelihood ridges and multimodality in population growth rate models. Ecology. 2009; 90(8):2313–2320. https://doi.org/10.1890/08-1461.1 PMID: 19739392

51. Swendsen RH, Wang JS. Replica Monte Carlo simulation of spin-glasses. Phys Rev Lett. 1986; 57(21):2607. https://doi.org/10.1103/PhysRevLett.57.2607 PMID: 10033814

52. Kou S, Zhou Q, Wong WH. Equi-energy sampler with applications in statistical inference and statistical mechanics. Ann Stat. 2006; 34(4):1581–1619. https://doi.org/10.1214/009053606000000515

53. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995; 82(4):711–732. https://doi.org/10.1093/biomet/82.4.711

54. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005; 14(8):2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x PMID: 15969739

55. Parry HR. Cereal aphid movement: general principles and simulation modelling. Mov Ecol. 2013; 1:14. https://doi.org/10.1186/2051-3933-1-14 PMID: 25709827

56. Parry HR, Evans AJ, Morgan D. Aphid population response to agricultural landscape change: a spatially explicit, individual-based model. Ecol Model. 2006; 199(4):451–463. https://doi.org/10.1016/j.ecolmodel.2006.01.006

57. Dallot S, Gottwald T, Labonne G, Quiot JB. Spatial pattern analysis of sharka disease (*Plum pox virus* strain M) in peach orchards of southern France. Phytopathology. 2003; 93:1543–1552. https://doi.org/10.1094/PHYTO.2003.93.12.1543 PMID: 18943618

58. Dallot S, Gottwald T, Labonne G, Quiot JB. Factors affecting the spread of *Plum pox virus* strain M in peach orchards subjected to roguing in France. Phytopathology. 2004; 94:1390–1398. https://doi.org/10.1094/PHYTO.2004.94.12.1390 PMID: 18943711

59. Gottwald T, Avinent L, Llácer G, Hermoso de Mendoza A, Cambra M. Analysis of the spatial spread of sharka (plum pox virus) in apricot and peach orchards in eastern Spain. Plant Dis. 1995; 79(3):266–278. https://doi.org/10.1094/PD-79-0266

Text A: Probabilistic Framework for Statistical Inference

Epidemic reconstruction is performed in a Bayesian framework, using one time step per year for estimation. For a fixed number κ of introduction patches, estimation is based on the posterior distribution of parameter set Θ given observed data Y:

$$f(\Theta|Y,\kappa) = \frac{\widehat{l}(Y|\Theta)f(\Theta|\kappa)}{\int \widehat{l}(Y|\Theta)f(\Theta|\kappa)d\Theta}, \tag{S1}$$

where $f(\Theta|Y,\kappa)$ is the joint probability density of parameters Θ given data Y and κ, $f(\Theta|\kappa)$ is the prior density of Θ given κ, and $\widehat{l}(Y|\Theta)$ is the pseudo-likelihood of Y given Θ. We call $\widehat{l}$ a pseudo-likelihood since it exploits certain approximations to simplify computation (see below). Here, Y and Θ simply represent "all data" and "all parameters" respectively; a more detailed presentation is given below (Texts B and C in S1 Texts).

In what follows, $T_{E,i}$, $T_{H,i}$, $T_{D,i}$ and $T_{R,i}$ are the dates at which a given (randomly selected) individual from patch i makes the transition to compartments E, H, D and R respectively. For a given individual of patch i, $T_{D,i}$ is the date at which symptoms were observed (for the first time), so $T_{D,i}$ corresponds either to one of the K_i inspection dates in patch i, or is right-censored. Since our model is designed for the analysis of patch-level data, we assume that $p(T_{D,i} = t_{i,k})$ (the probability for a tree to be detected at the k^{th} inspection date in patch i) is homogeneous within patch i at any given time. Let N_i be the initial number of trees planted in patch i, $D^+_{i,k}$ be the number of newly detected infectious individuals (among symptomatic individuals) at each inspection k, and D^-_i be the number of (uninfected or infected) individuals upon which symptoms were not detected in any of the K_i inspections. Based on Eq (12), and noting that $D^-_i = N_i - \sum_{k=1}^{K_i} D^+_{i,k}$, the pseudo-likelihood of the observed data is:

$$\widehat{l}(Y|\Theta) = \prod_i M_i \left[\left(1 - \sum_{k=1}^{K_i} p(T_{D,i} = t_{i,k})\right)^{N_i - \sum_{k=1}^{K_i} D^+_{i,k}} \prod_{k=1}^{K_i} p(T_{D,i} = t_{i,k})^{D^+_{i,k}} \right], \tag{S2}$$

where the multinomial coefficient for patch i, M_i, is constant with respect to Θ. We now address each transition in the $SEHDR$ framework to indicate how spatial structure and transition-time censoring are accounted for when constructing the probabilities in Eq (S2).

T_E: Transmission

Let r denote the time step of the model. The probability that a given individual in patch i became infected in time interval $(t_{r-1}, t_r]$ is modelled as:

$$p(t_{r-1} < T_{E,i} \leq t_r) = p(t_{r-1} < T_{E,i} \leq t_r | t_{r-1} < T_{E,i}) \times p(t_{r-1} < T_{E,i}). \tag{S3}$$

Let $T_{0,i}$ denote the time at which patch i is initialised (i.e. the beginning of the first year of budbreak in patch i). Let Z_i be a Boolean random variable indicating whether (Z_i=1) or not (Z_i=0) patch i is an introduction patch (with initial prevalence p_i). The probability that an individual is still susceptible at t_{r-1} is:

$$p(t_{r-1} < T_{E,i}) = p(T_{0,i} < T_{E,i} | Z_i) \prod_{r'=1}^{r-1} p(t_{r'} < T_{E,i} | t_{r'-1} < T_{E,i}), \tag{S4}$$

where $p(T_{0,i} < T_{E,i}|Z_i) = 1 - p_i Z_i$, and where the conditional probability of remaining susceptible throughout a time step is:

$$p(t_r < T_{E,i} | t_{r-1} < T_{E,i}) = 1 - p(t_{r-1} < T_{E,i} \leq t_r | t_{r-1} < T_{E,i}). \tag{S5}$$

Assuming independence between infection events within each time interval $(t_{r-1}, t_r]$, the probability that at least one infection event affected an individual within $(t_{r-1}, t_r]$ is:

$$p(t_{r-1} < T_{E,i} \leq t_r | t_{r-1} < T_{E,i}) = 1 - e^{-\lambda_{i,t_r}}, \tag{S6}$$

and thus:

$$p(t_{r-1} < T_{E,i} \leq t_r) = (1 - p_i Z_i)(1 - e^{-\lambda_{i,t_r}}) e^{-\sum_{r'=1}^{r-1} \lambda_{i,t_{r'}}}, \tag{S7}$$

where the force of infection λ_{i,t_r} is defined (main text, Eq (5)) as the expected number of infection events affecting a given individual of patch i in $(t_{r-1}, t_r]$. Note that $I_{i,t_{r-1}}$ and $R_{i,t_{r-1}}$ in Eq (5) are unknown. Instead of using data augmentation to directly incorporate hundreds of unobserved infection times within the MCMC algorithm, we used a pseudo-likelihood [1] where these variables are replaced by their expected values:

$$E[I_{i,t_{r-1}}] = N_i \times \left[1 - p(t_{r-1} < T_{H,i}) - p(T_{R,i} \leq t_{r-1})\right] \text{ and} \tag{S8}$$

$$E[R_{i,t_{r-1}}] = N_i \times p(T_{R,i} \leq t_{r-1}). \tag{S9}$$

The term in square brackets in Eq (S8) gives the expected proportion of individuals in state H or D at time t_{r-1}. Similarly, the probability term in Eq (S9) gives the expected proportion of individuals in state R at time t_{r-1}. The terms $p(t_{r-1} < T_{H,i})$ and $p(T_{R,i} \leq t_{r-1})$ are derived below (in Eqs (S12) and (S17), respectively).

T_H: the end of the latent period

The marginal probability for an individual to be in compartment S or E at time t_r is found by integrating over all possibilities of the unknown infection time $T_{E,i}$ as follows:

$$\begin{aligned} p(t_r < T_{H,i}) =& p(t_r < T_{E,i}) + p(t_r < T_{H,i} | T_{E,i} \leq T_{0,i}) \times p(T_{E,i} \leq T_{0,i}) + \\ & \sum_{r'=1}^{r} \Big[p(t_r < T_{H,i} | t_{r'-1} < T_{E,i} \leq t_{r'}) \times p(t_{r'-1} < T_{E,i} \leq t_{r'}) \Big]. \end{aligned} \tag{S10}$$

The conditional probability for an individual to be still in the exposed state at t_r given that it became infected in a previous time interval $(t_{r'-1}, t_{r'}]$ is modelled as:

$$p(t_r < T_{H,i} | t_{r'-1} < T_{E,i} \leq t_{r'}) = 1 - F_{\text{Tr,lat}}(t_r - t_{r'}), \tag{S11}$$

where $F_{\text{Tr,lat}}$ is the cumulative distribution function of the left-truncated gamma distribution with shape θ_1 and scale θ_2. In the case of PPV-M the truncation is used to represent a minimal latent period of one winter, and we assume $p(t_r < T_{H,i} | T_{E,i} \leq T_{0,i}) = 0$. Thus:

$$p(t_r < T_{H,i}) = p(t_r < T_{E,i}) + \sum_{r'=1}^{r} \Big(\left[1 - F_{\text{Tr,lat}}(t_r - t_{r'})\right] \times p(t_{r'-1} < T_{E,i} \leq t_{r'}) \Big). \tag{S12}$$

T_D: Detection

At each inspection, infectious trees can be detected on the basis of sharka-specific symptoms that appear mainly on flowers and leaves when the latent period is over. Thus, we assume that only trees in state H can be detected (as infected). The probability that symptoms are detected for the first time on a given tree in patch i at inspection date $t_{i,k}$ is:

$$\begin{aligned} p(T_{D,i} = t_{i,k}) &= p(T_{D,i} = t_{i,k}, T_{D,i} > t_{i,k-1}, T_{H,i} \leq t_{i,k}) \\ &= p(T_{D,i} = t_{i,k} | T_{D,i} > t_{i,k-1}, T_{H,i} \leq t_{i,k}) \times p(T_{D,i} > t_{i,k-1}, T_{H,i} \leq t_{i,k}), \end{aligned} \tag{S13}$$

where the first term corresponds to the detection sensitivity ρ_{i,t_r}, and the second term is derived as follows:

$$\begin{aligned} p(T_{D,i} > t_{i,k-1}, T_{H,i} \leq t_{i,k}) &= p(T_{H,i} \leq t_{i,k}) - p(T_{D,i} \leq t_{i,k-1}, T_{H,i} \leq t_{i,k}) \\ &= p(T_{H,i} \leq t_{i,k}) - p(T_{D,i} \leq t_{i,k-1}) \times p(T_{H,i} \leq t_{i,k} | T_{D,i} \leq t_{i,k-1}). \end{aligned} \tag{S14}$$

Because $T_{H,i} \leq T_{D,i}$ for any given individual, the last term of Eq (S14) is equal to one. Thus, we obtain the relation:

$$p(T_{D,i} > t_{i,k-1}, T_{H,i} \leq t_{i,k}) = p(T_{H,i} \leq t_{i,k}) - \sum_{k'<k} p(T_{D,i} = t_{i,k'}). \tag{S15}$$

The probability $p(T_{H,i} \leq t_{i,k})$ of having passed into compartment H prior to, or at, $t_{i,k}$ is derived directly from Eq (S12). Note that when more than one inspection date falls within a single time step, the order of these inspections is preserved, i.e. Eqs (S13) and (S15) do not change and the probability of detecting an infectious tree for the first time at the second inspection date will be lower than at the first date.

T_R: Removal

Although removal dates for whole orchards were recorded, this was not the case for the removal of individual trees. We adopt a discrete-time survival model to account for this censoring in which detected infectious trees in patch i are removed each time step with a probability determined by the mean duration δ between detection and removal. Thus, the probability for a tree in patch i to be removed before time t_r is modelled as:

$$p(T_{R,i} < t_r) = \sum_{\{k:\, t_{i,k} < t_r\}} \Big[p(T_{R,i} < t_r | T_{D,i} = t_{i,k}) \times p(T_{D,i} = t_{i,k}) \Big] \tag{S16}$$

$$= \sum_{\{k:\, t_{i,k} < t_r\}} \Big(\Big[1 - p(T_{R,i} > t_r | T_{R,i} > t_{r-1}, T_{D,i} = t_{i,k})^{\Delta(t_r - t_{i,k})} \Big] \times p(T_{D,i} = t_{i,k}) \Big), \tag{S17}$$

where $\Delta(t_r - t_{i,k})$ is the number of time steps between t_r and the start of the time step containing inspection date $t_{i,k}$. Here, $p(T_{R,i} > t_r | T_{R,i} > t_{r-1}, T_{D,i} = t_{i,k})$ is the probability that an individual in state D at time t_{r-1} still is in state D by time t_r. In practice, the vast majority of detected trees are removed within the legal 10-day delay, and almost all detected trees are removed before the end of the growing season. However, since we use a 1-year time step for estimation, we assume $p(T_{R,i} > t_r | T_{R,i} > t_{r-1}, T_{D,i} = t_{i,k})$ is zero when t_r and $t_{i,k}$ belong to different civil years and is one otherwise.

Text B: Prior Distributions

The priors used throughout this work are:

$$\begin{aligned}
\mu &\sim \text{Uniform}(0, 1), \\
\sigma &\sim \text{Exponential}(\text{scale} = 10^3), \\
\theta_1 &\sim \text{Gamma}(15.5, 0.444), \\
\theta_2 &\sim \text{Gamma}(5.11, 0.888), \\
\log(\beta) &\sim \text{Uniform}(-\infty, \infty), \\
\rho &\sim \text{Beta}(559, 141), \\
Z_i &\sim \text{Bernoulli}(0.5), \\
p_i | Z_i = 1 &\sim \text{Uniform}(0, 1), \\
\widehat{X}^{\text{mis}}_{\text{pl},i} &\sim \text{Empirical}(\boldsymbol{X}_{\text{pl}}), \\
\widehat{X}^{\text{mis}}_{\text{in},i,k} | \mathcal{T}_{i,k} &\sim \text{Empirical}(\boldsymbol{X}_{\text{in}} | \mathcal{T}_{i,k}),
\end{aligned}$$

where: $\mu = \frac{s_1}{s_1+s_2}$ and $\sigma = s_1 + s_2$ are parameters of the BWME kernel (Eq 11); θ_1 and θ_2, respectively, are shape and scale parameters for the latent period model (Eq 2); β is the transmission coefficient (Eq 5); ρ is the detection sensitivity (Eq 3); Z_i is a variable indicating whether patch i is an introduction patch (Z_i=1) or not (Z_i=0); p_i, the introduction prevalence in patch i (Eqs S4 and S7), is set to zero when patch i is not an introduction patch; $\widehat{X}^{\text{mis}}_{\text{pl},i}$ are imputed values for missing planting dates which are assumed *a priori* to be distributed according to the empirical distribution of known planting dates $\boldsymbol{X}_{\text{pl}}$; $\widehat{X}^{\text{mis}}_{\text{in},i}$ are imputed values for missing inspection dates which are assumed *a priori* to be distributed according to the empirical distribution of known inspection dates $\boldsymbol{X}_{\text{in}}$ conditioned on inspection type data $\mathcal{T}_{i,k}$ indicating whether inspection k in patch i focused on symptoms on either flowers (i.e. early spring) or leaves (i.e. late spring). Based on field and laboratory observations by SD and GL, hyper-parameters for the latent period model are set using weeks as the temporal unit, with a prior mean and variance of 6.9 and 3.1 for θ_1, and 4.5 and 4.0 for θ_2 (which corresponds to θ_{exp}=31 and θ_{var}=435). Hyper-parameters for the prior sensitivity to detect infectious individuals are based on available field data and give a mode et 0.80 and a variance of 0.00023.

For the study of the impact of ρ on estimation, the prior distribution of detection sensitivity is defined as Beta(1+$\rho\omega$, 1+(1-ρ)ω), with ω=1, 100 or 10000 corresponding to weak, mild or strong prior knowledge, respectively. The mode of these priors matches the simulated value of ρ and the associated precision increases with ω (e.g. for ρ=0.8, the variance of the prior is equal to 0.06, 1.6×10^{-3}, and 1.6×10^{-6}, respectively).

Text C: Markov Chain Monte Carlo

Bayesian inference is based on Markov chain Monte Carlo (MCMC) approximation of the joint distribution:

$$f(\mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}, \boldsymbol{Z}, \widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}}, \widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}} \,|\, \boldsymbol{Y}, \boldsymbol{X}_{\text{pl}}, \boldsymbol{X}_{\text{in}}, \boldsymbol{\mathcal{T}}, \kappa), \tag{S18}$$

where: $f(\cdot|\cdot)$ represents the conditional probability density of a given subset of parameters; $\boldsymbol{Z}=\{Z_1, \ldots, Z_i, \ldots\}$ is the set of variables specifying whether each patch i is classified as an introduction patch (Z_i=1) or not (Z_i=0); $\boldsymbol{p}=\{p_1, \ldots, p_i, \ldots\}$ is the corresponding set of introduction prevalences; $\boldsymbol{Y}$ represents the data associated with each inspection, including the number of infectious individuals detected, for the first time, in each orchard at each inspection; $\boldsymbol{X}_{\text{pl}}$ and $\boldsymbol{X}_{\text{in}}$ are sets of known planting and inspection dates respectively; $\widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}}$ and $\widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}}$ are sets of imputed values for unknown planting and inspection dates; and $\boldsymbol{\mathcal{T}}$ is the set of inspection type data (inspection on flowers or leaves). Thus, Eq (S18) provides the full form of the posterior distribution written in a simplified form in Eq (S1).

A Gibbs sampler is used to sequentially (in random order) sample subsets of parameters using the following set of conditional distributions:

$$f(\boldsymbol{Z} \,|\, \mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}, \widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}}, \widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}}, \boldsymbol{Y}, \boldsymbol{X}_{\text{pl}}, \boldsymbol{X}_{\text{in}}, \kappa), \tag{S19}$$

$$f(\widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}} \,|\, \mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}, \boldsymbol{Z}, \widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}}, \boldsymbol{Y}, \boldsymbol{X}_{\text{pl}}, \boldsymbol{X}_{\text{in}}, \kappa), \tag{S20}$$

$$f(\widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}} \,|\, \mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}, \boldsymbol{Z}, \widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}}, \boldsymbol{Y}, \boldsymbol{X}_{\text{pl}}, \boldsymbol{X}_{\text{in}}, \boldsymbol{\mathcal{T}}, \kappa), \tag{S21}$$

$$f(\mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}, \,|\, \boldsymbol{Z}, \widehat{\boldsymbol{X}}_{\text{pl}}^{\text{mis}}, \widehat{\boldsymbol{X}}_{\text{in}}^{\text{mis}}, \boldsymbol{Y}, \boldsymbol{X}_{\text{pl}}, \boldsymbol{X}_{\text{in}}, \kappa). \tag{S22}$$

To ensure initialisation of the MCMC with a classification $\boldsymbol{Z}$ generating a finite log-likelihood, each chain is initialised with a number of introduction patches κ' much larger than the imposed number κ. The extra introduction patches are removed using greedy Metropolis-Hastings steps until $\kappa'=\kappa$ (in practice, this takes just a few hundreds of iterations). Thereafter, the conditional distribution for the set $\boldsymbol{Z}$ (Eq S19) is sampled using block proposals (that maintain a constant κ) of various sizes in a Metropolis-Hastings sampler. Data augmentation steps (Eqs S20 and S21) are performed using block Metropolis-Hastings samplers, where block proposals of various sizes are drawn from the empirical distribution associated with each missing data component.

The distribution depicted in Eq (S22) is sampled as follows. Let $\widetilde{\Theta}$ represent a transformation of the parameter block $\Theta=\{\mu, \sigma, \beta, \theta_1, \theta_2, \rho, \boldsymbol{p}\}$ that enables to sample these parameters on an unbounded parameter space. In $\widetilde{\Theta}$, parameters ρ and p_i that are defined on (0,1) are sampled on the logit scale, and parameters that are not defined below zero ($\mu, \sigma, \theta_1, \theta_2, \beta$) are sampled on the logarithmic scale. Where appropriate, standard change-of-variable corrections are used to transform prior distributions to their associated sampling scales. The transformed parameter block is sampled using a Metropolis-Hastings algorithm [2] with multivariate Gaussian proposal distribution and adaptive covariance matrix. The details of this adaptive scheme are given below.

Various adaptive Metropolis-Hastings algorithms have been proposed [3–5]. We use a standard approach with proposals $\widetilde{\Theta}_t^*$ generated from:

$$\widetilde{\Theta}_t^* \sim \mathcal{N}\Big(\widetilde{\Theta}_{t-1}, \frac{2.38^2}{\dim(\widetilde{\Theta}_{t-1})}\widehat{\Sigma}_{t-1}\Big), \tag{S23}$$

where $\widehat{\Sigma}_{t-1}$ is an estimate of covariance in the posterior samples of $\widetilde{\Theta}$. This scheme is known to provide optimal mixing in simple theoretical examples [6]. Estimates of the covariance in Eq (S23) are updated during adaptive burn-in using the following iterative procedure:

$$\widehat{E}[\widetilde{\Theta}_{t+1}] = \frac{(\mathcal{C}_t - 1)\widehat{E}[\widetilde{\Theta}_t] + \widetilde{\Theta}_t}{\mathcal{C}_t} \tag{S24}$$

$$\Delta_t = \widetilde{\Theta}_t - \widehat{E}[\widetilde{\Theta}_{t+1}]$$

$$\widehat{\Sigma}_{t+1} = \frac{\Delta_t \otimes \Delta_t}{\mathcal{C}_t} + \frac{\mathcal{C}_t - 2}{\mathcal{C}_t - 1}\widehat{\Sigma}_t \tag{S25}$$

where $\widehat{E}[\widetilde{\Theta}_{t+1}]$ is an estimate of the sample mean at iteration t+1, $\otimes$ is the outer product, and $\mathcal{C}_t$ is a counter. In many adaptive Metropolis-Hastings samplers, $\mathcal{C}_t$ is simply set to the number of iterations of the MCMC.

However, such schemes can become prematurely inflexible, and this can lead to suboptimal mixing. To avoid this problem, we use a counter that only increases when the sampler accepts a proposal and decreases each time the sampler encounters a new area of parameter space that provides a non-negligible increase in the maximum posterior log-likelihood estimate. Thus, we define counter $\mathcal{C}_t$ as follows:

$$L'_t = \max(L_t, L'_{t-1}) \tag{S26}$$
$$\mathcal{Z}_t = \mathcal{Z}_{t-1} \exp(L'_{t-1} - L'_t) + \exp(L_t - L'_t) \tag{S27}$$
$$\mathcal{W}_t = \exp(L_t - L'_t)/\mathcal{Z}_t \tag{S28}$$
$$\mathcal{C}_t = \max\Big(\mathcal{C}_{\min}, \mathbb{1}_{\text{accept}(\widetilde{\Theta}^*_t)} + (1 - \mathcal{W}_t) \times \mathcal{C}_{t-1}\Big), \tag{S29}$$

where L_t is the log-likelihood at the end of iteration t and L'_t indicates the maximal log-likelihood encountered since the first iteration. When a large gain in L' is encountered, $\mathcal{W}_t \approx 1$; this effectively resets $\mathcal{C}_t$ to some specified minimal value $\mathcal{C}_{\min}$ (we use $\mathcal{C}_{\min}$=250). Conversely, when no (or only negligible) augmentation of L' has been encountered for a long time, $\mathcal{Z}_t$ is free to grow (we use $\mathcal{Z}_0$=1), the weights $\mathcal{W}_t$ become small, and $\mathcal{C}_t$ increases by approximately one unit each time the sampler (Eq S22) accepts a proposal. Clearly, once the MCMC has generated a sample very close to the maximum posterior log-likelihood, $\mathcal{C}_t$ starts to grow monotonically and then the adaptation scheme becomes less and less flexible. Adaptation of the covariance in Eq (S25) is used during a burn-in period that is stopped once $\mathcal{C}_t > \mathcal{C}_{\text{Target}} + \mathcal{C}_{\min}$. Adaptive burn-in is turned back on if either i) further sampling generates large log-likelihood gains that reduce $\mathcal{C}_t$ sufficiently to violate this inequality, or ii) the acceptance rate becomes less than 1/200. In case (ii), we reset $\mathcal{C}_t$ to $\mathcal{C}_{\min}$. We set $\mathcal{C}_{\text{Target}}$=10000 for the simulation studies and $\mathcal{C}_{\text{Target}}$=20000 for the analysis of the real epidemic.

Two constraints are employed to prevent convergence to degenerate solutions in early MCMC iterations: i) any Metropolis-Hastings proposal giving a mean latent period greater than 10 years is rejected; ii) Metropolis-Hastings proposals are rejected if the expected prevalence is greater than 30% following an uncontrolled 4-year epidemic initialised with a single infectious individual in a landscape comprising just one patch. Constraint (ii) involves simulating an SEH sub-model in the patch containing the median number of planted trees. Since the areas of parameter space banned by these constraints are extremely unrealistic for our biological model, any bias associated with these truncations is assumed to be negligible. Indeed, analysis of MCMC output indicates that posterior distributions are located far from the bounds imposed on the parameter space.

In the simulation studies, 10 chains were run for 25000 post burn-in iterations for each simulated epidemic. To analyse the real epidemic, 30 chains were run for 10^5 post burn-in iterations for each value of κ. In both cases, the sampled parameters were stored every 25 iterations. Unless stated otherwise, all results are based on the chain with the highest mean posterior likelihood. Simulation and estimation algorithms are written in C and called from R. MCMC convergence diagnostics are performed using the R package *coda* [7]. Calculation of Eq (6) of the main text is performed using the DCUTRI algorithm of the software CaliFloPP [8].

Text D: Model Selection for K

To identify the number of introduction patches κ, we use the Fisher information of the sample $\mathcal{I}(\kappa)$, which is inversely proportional to the variance of the estimator of κ. This relationship implies that low-variance estimators can be found at the point of greatest curvature in the log-likelihood of the observed data. Thus, we seek the value of κ that maximises the expectation:

$$\mathcal{I}(\kappa) = E\left[-\Delta^2(L_{Y;\kappa})|\Theta\right] \tag{S30}$$
$$= \int -\Delta^2(L_{y;\kappa}) l(y;\kappa) dy \tag{S31}$$
$$\approx -\frac{1}{C}\sum_{c=1}^{C}\left(L_{\widehat{Y}_c;\kappa+1} - 2L_{\widehat{Y}_c;\kappa} + L_{\widehat{Y}_c;\kappa-1}\right) \tag{S32}$$
$$\approx -\bar{L}_{Y,\kappa+1} + 2\bar{L}_{Y,\kappa} - \bar{L}_{Y,\kappa-1}, \tag{S33}$$

where Δ^2 is the centred second difference operator and $L_{Y;\kappa} = \ln\left[l(Y|\Theta,\kappa)\right]$ is the logarithm of the observed data likelihood. One possible approximation of the integral (Eq S31) is to use Monte Carlo approximation (Eq

S32) where $\widehat{Y}_c$ are replicated datasets generated during the MCMC runs for κ+1, κ and κ-1. An alternative approximation (Eq S33) is the second difference in the mean log-likelihoods of the observed data generated by the MCMC, given κ. We use this second approach since it has the advantage over Eq (S32) that no computation time is spent performing additional simulations. Approaches of this kind lead to estimators that are robust against over-parametrisation [9].

References

1. Gourieroux C, Monfort A, Trognon A. Pseudo Maximum Likelihood Methods: Theory. Econometrica. 1984;52(3):681–700.
2. Hastings WK. Monte Carlo sampling methods using Markov chains and their applications. Biometrika. 1970;57(1):97–109.
3. Haario H, Saksman E, Tamminen J. An adaptive Metropolis algorithm. Bernoulli. 2001;7(2):223–242.
4. Vihola M. Robust adaptive Metropolis algorithm with coerced acceptance rate. Stat Comput. 2012;22(5):997–1008.
5. Griffin JE, Walker SG. On adaptive Metropolis-Hastings methods. Stat Comput. 2013;23(1):123–134.
6. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian Data Analysis. 2nd ed. Texts in Statistical Science Series. Chapman & Hall/CRC; 2004.
7. Plummer M, Best N, Cowles K, Vines K. CODA: Convergence diagnosis and output analysis for MCMC. R News. 2006;6(1):7–11.
8. Bouvier A, Kiêu K, Adamczyk K, Monod H. Computation of the integrated flow of particles between polygons. Environ Model Softw. 2009;24(7):843–849.
9. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14(8):2611–2620.

Supporting Figures

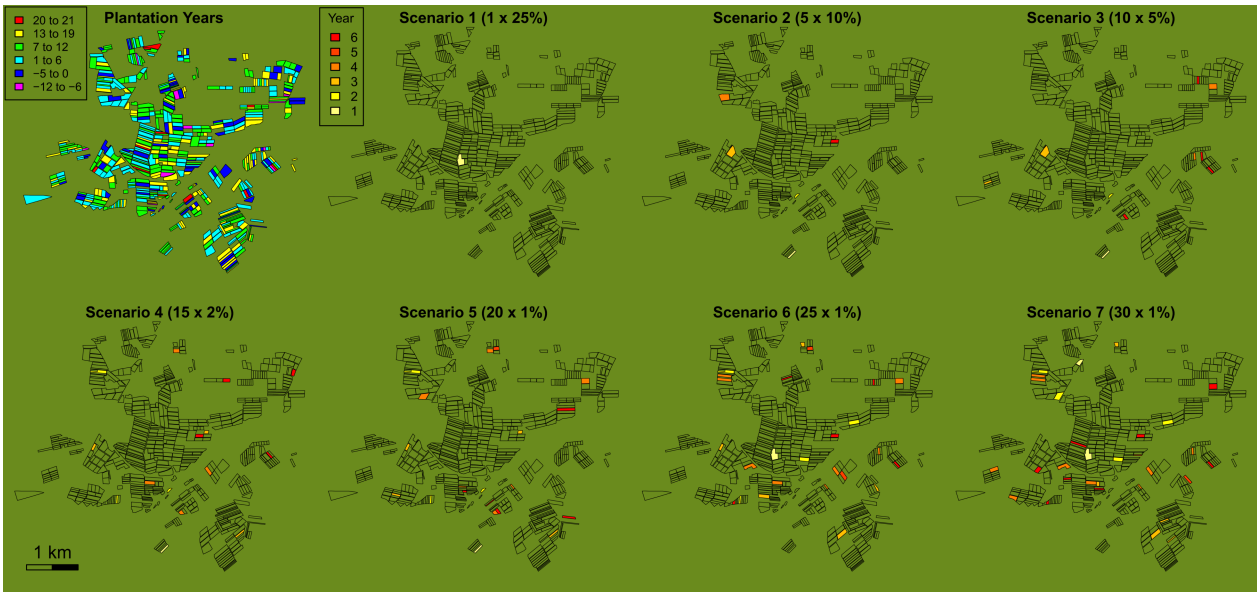

Fig S1. Simulated planting years and introduction patch locations used in the simulation study. The first map (top left) represents the randomisation of the first planting years of the 553 patches. These years were sampled without replacement from their empirical distribution. The other maps show the location and planting year of each introduction patch in the seven introduction scenarios. The number of introduction patches and their initial prevalence are indicated for each introduction scenario. Note the greater landscape connectivity in the central area.

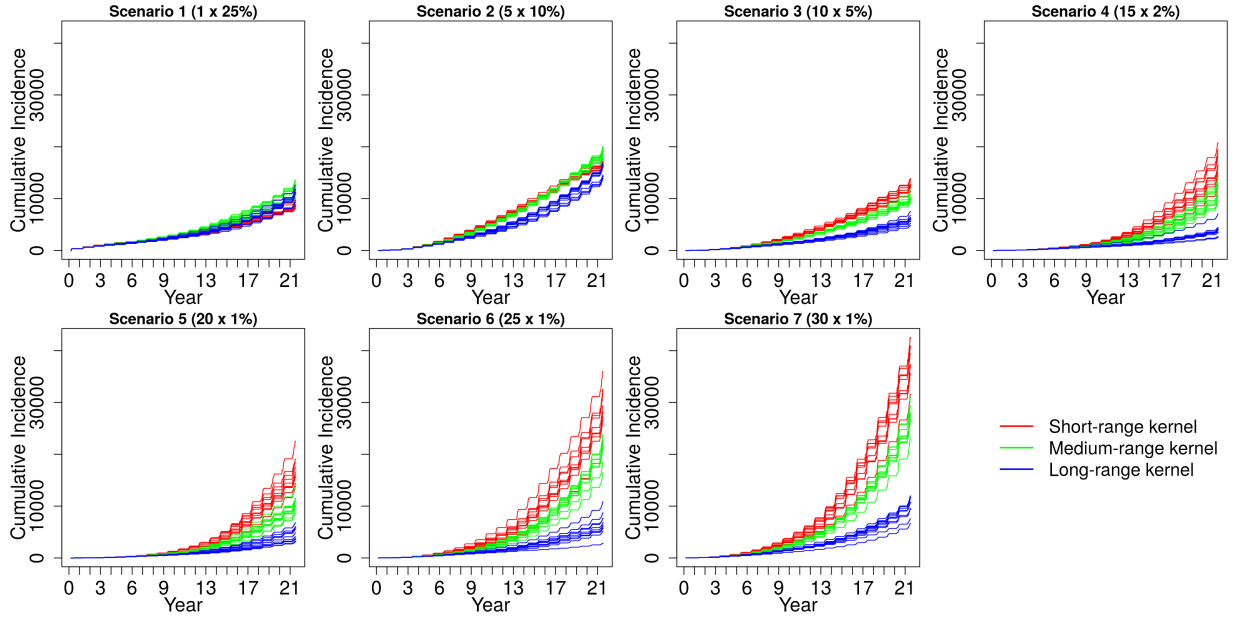

Fig S2. Cumulative detected incidence for the introduction scenarios tested in the simulation study. For each introduction scenario, the number of introduction patches and the corresponding initial disease prevalence are mentionned above the graph. The three tested kernels are represented by different colours. For each combination of kernel and introduction scenarios, 10 independant simulated epidemics are shown.

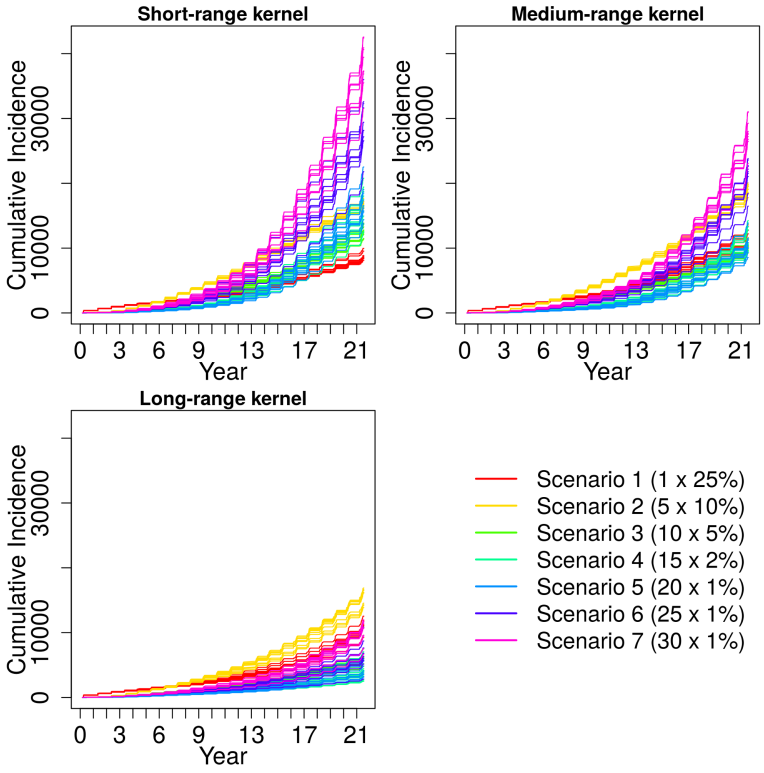

Fig S3. Cumulative detected incidence for the kernels tested in the simulation study. For each kernel, the seven tested introduction scenarios are represented by different colours. For each combination of kernel and introduction scenarios, 10 independant simulated epidemics are shown.

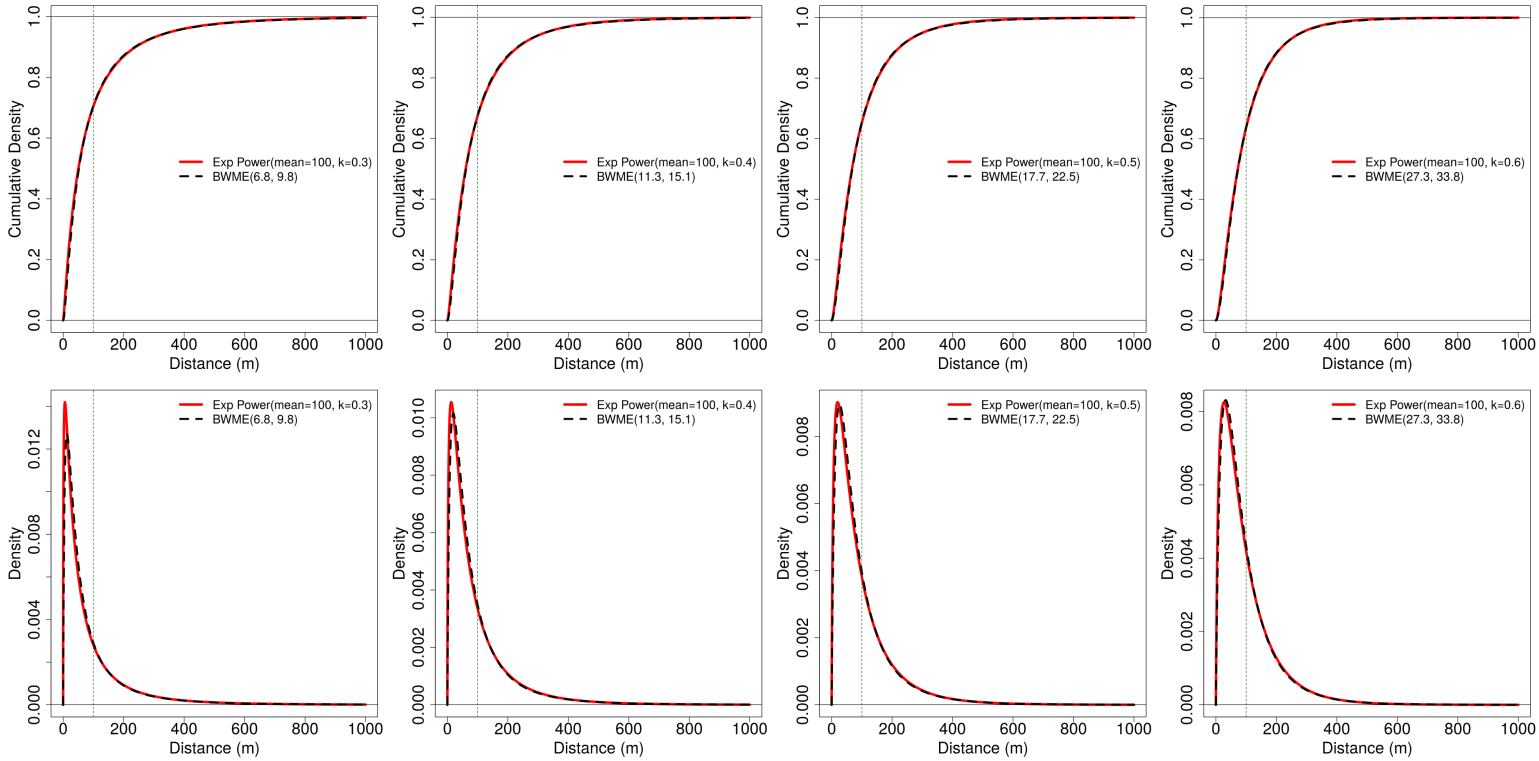

Fig S4. Best-fit BWME kernel approximations of exponential-power kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.

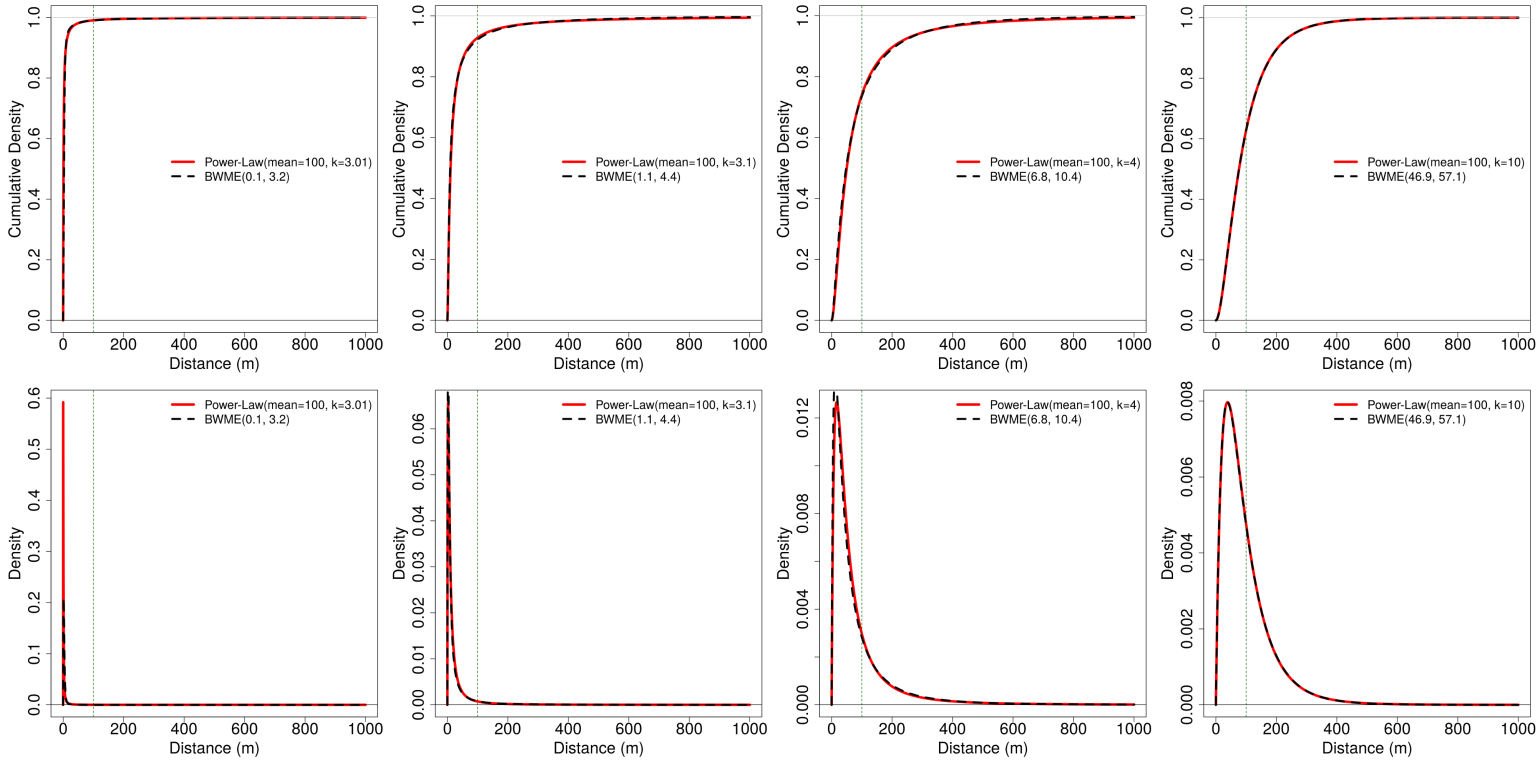

Fig S5. Best-fit BWME kernel approximations of power-law kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.

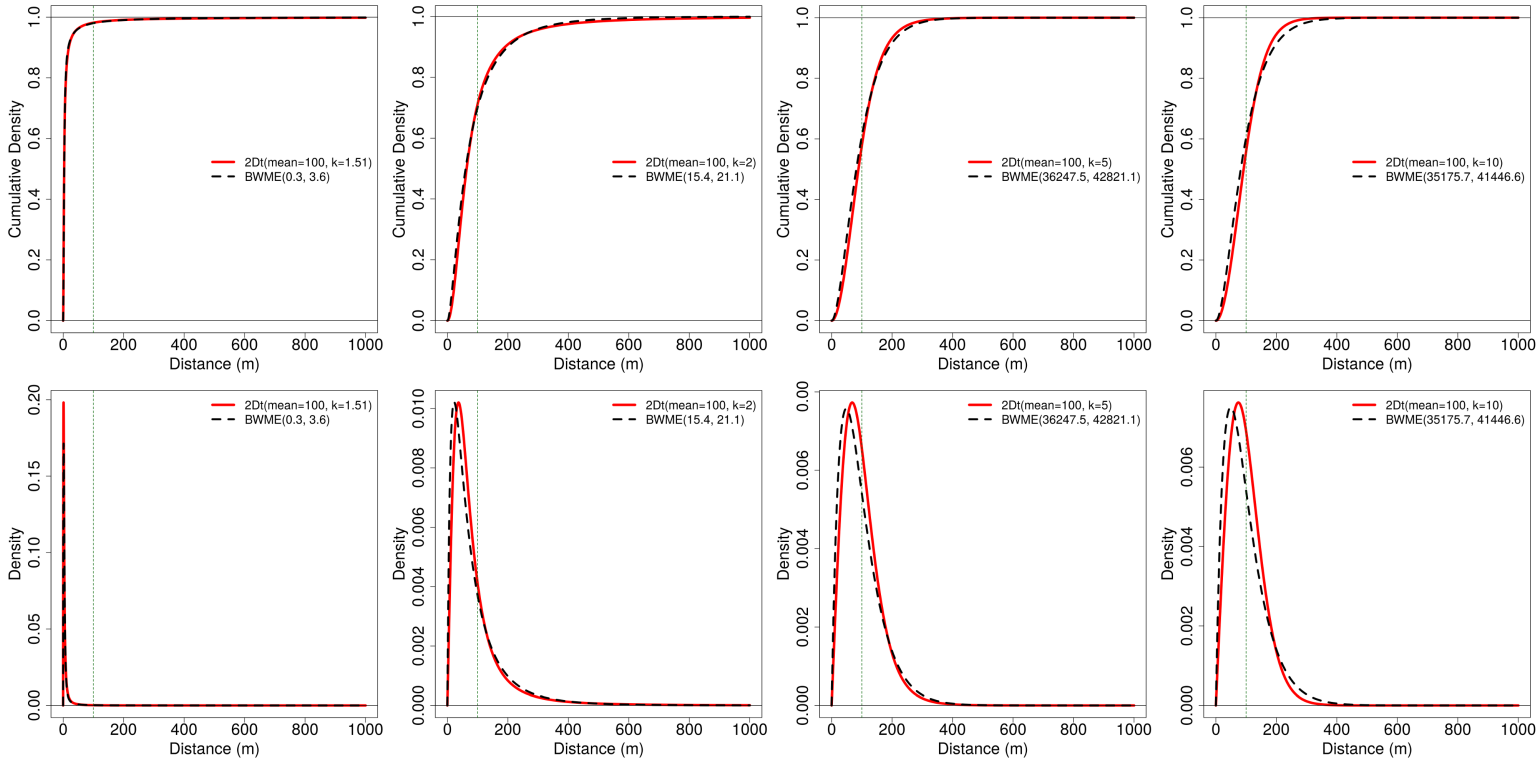

Fig S6. Best-fit BWME kernel approximations of 2Dt kernels. The kernels corresponding to 4 values of the shape parameter are represented by their cumulative distribution function F^{1D} (top) and the associated probability density function f^{1D} (bottom) of the distance travelled. Green dashed line: mean distance travelled.

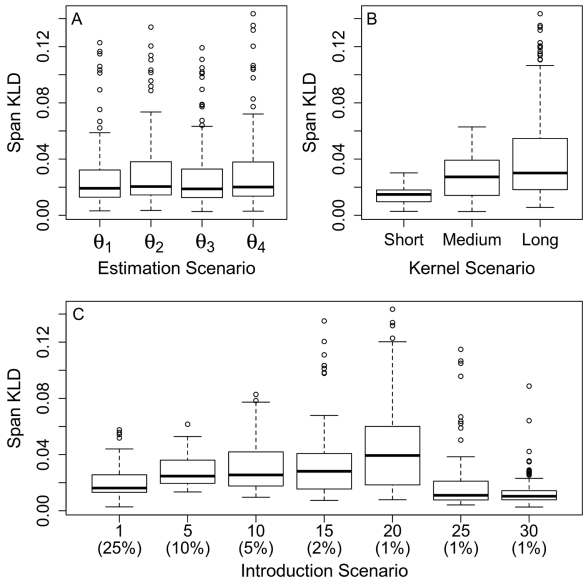

Fig S7. Boxplots of the variation among estimated dispersal kernels. Impact of (A) estimation scenario, (B) kernel range, and (C) disease introduction scenario [number of introduction patches (with initial disease prevalence)] on the precision of estimated dispersal kernels. Precision is measured by the span of the 95% credibility interval of Kullback-Leibler distances (Span KLD) between simulated and estimated dispersal kernels. Each panel consists of 840 points, which correspond to 10 epidemics × 7 disease introduction scenarios × 3 dispersal kernels × 4 parameter estimation schemes.

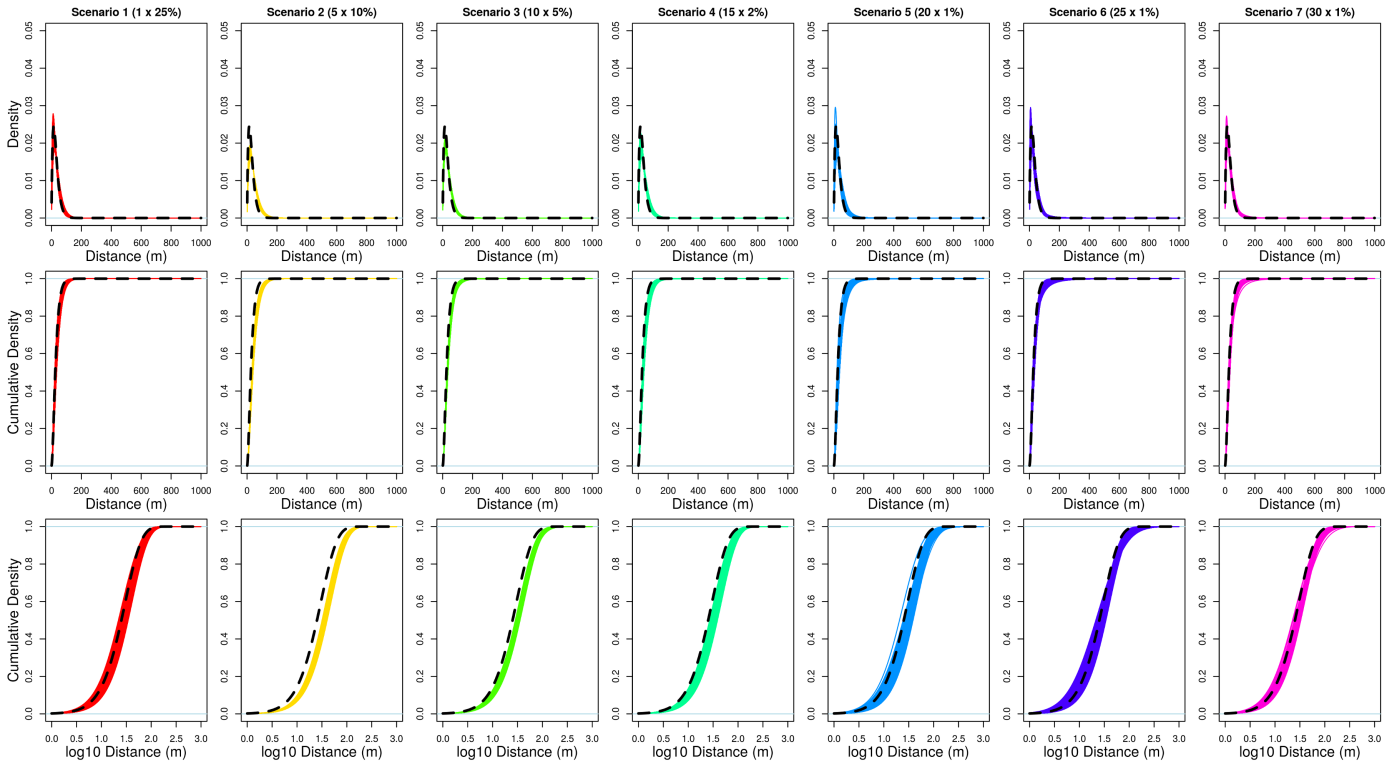

Fig S8. Influence of introduction scenarios on the estimation of a short-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a short-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood (mean proportion: 10%) increases quadratically with the number of source orchards. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).

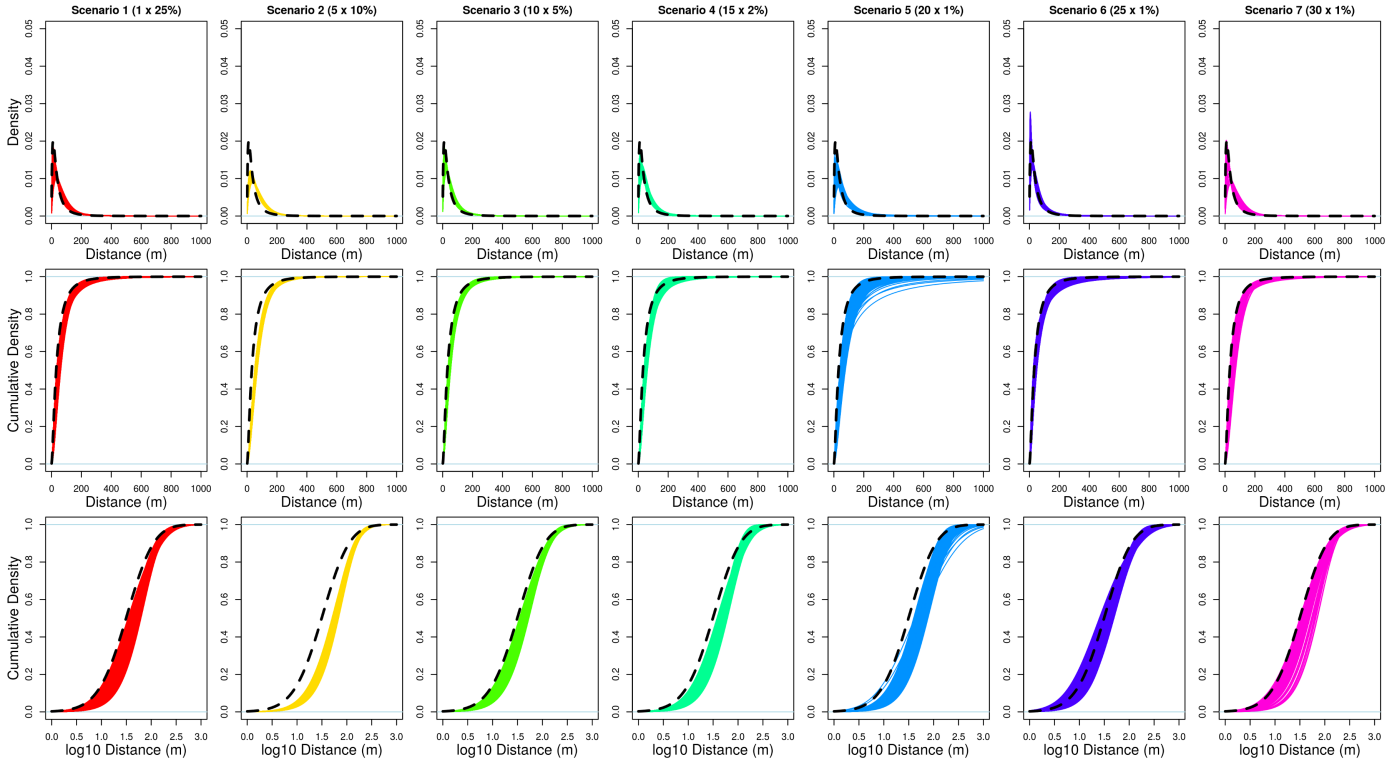

Fig S9. Influence of introduction scenarios on the estimation of a medium-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a medium-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood varies among introduction scenarios, with a mean proportion of 2.6%. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).

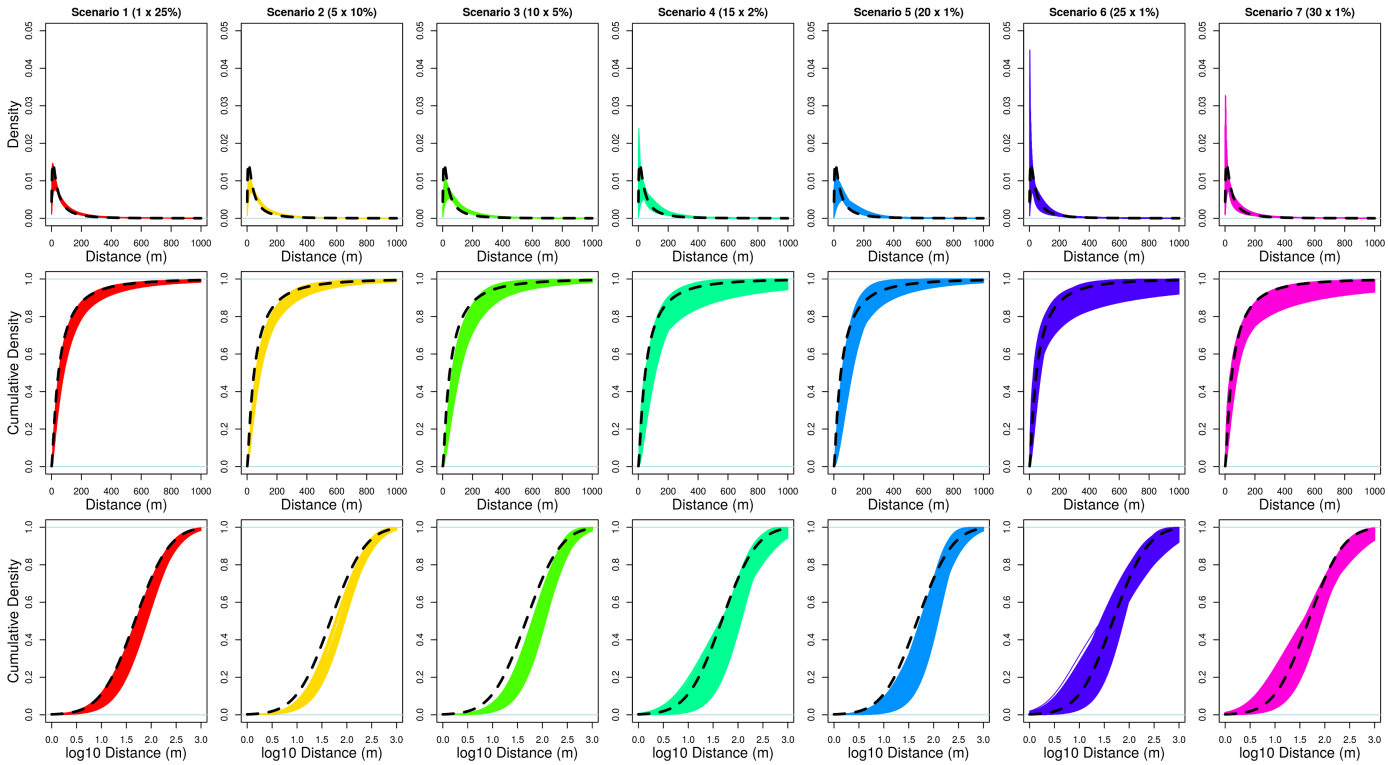

Fig S10. Influence of introduction scenarios on the estimation of a long-range dispersal kernel. For each introduction scenario, 10 epidemics were simulated with a long-range kernel (black dashed curve), and 10 MCMC chains were run per simulated epidemic. The posterior distributions of the kernel obtained under the most exhaustive estimation scheme (Θ_4) are represented for all chains with non-negligible mean posterior likelihood. The proportion of MCMC chains with negligible mean posterior likelihood is low (mean proportion: 0.4%) for all the introduction scenarios. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).

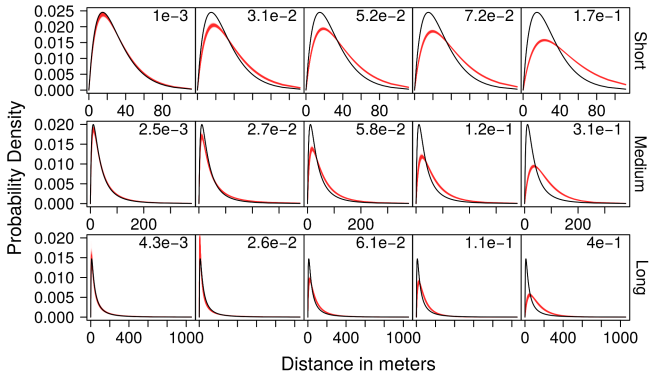

Fig S11. Comparison of simulated and estimated dispersal kernels. From left to right: kernels with the minimum, lower quartile, median, upper quartile and maximum Kullback-Leibler (KL) distances (posterior mean), for all chains with non-negligible mean posterior likelihood. Estimations (red) under the most exhaustive scheme (Θ_4) are based on simulated epidemics with short-, medium- and long-range kernels (from top to bottom; black). Kernels are represented by their marginal probability density function f^{1D}. The mean KL distance is indicated for each estimation.

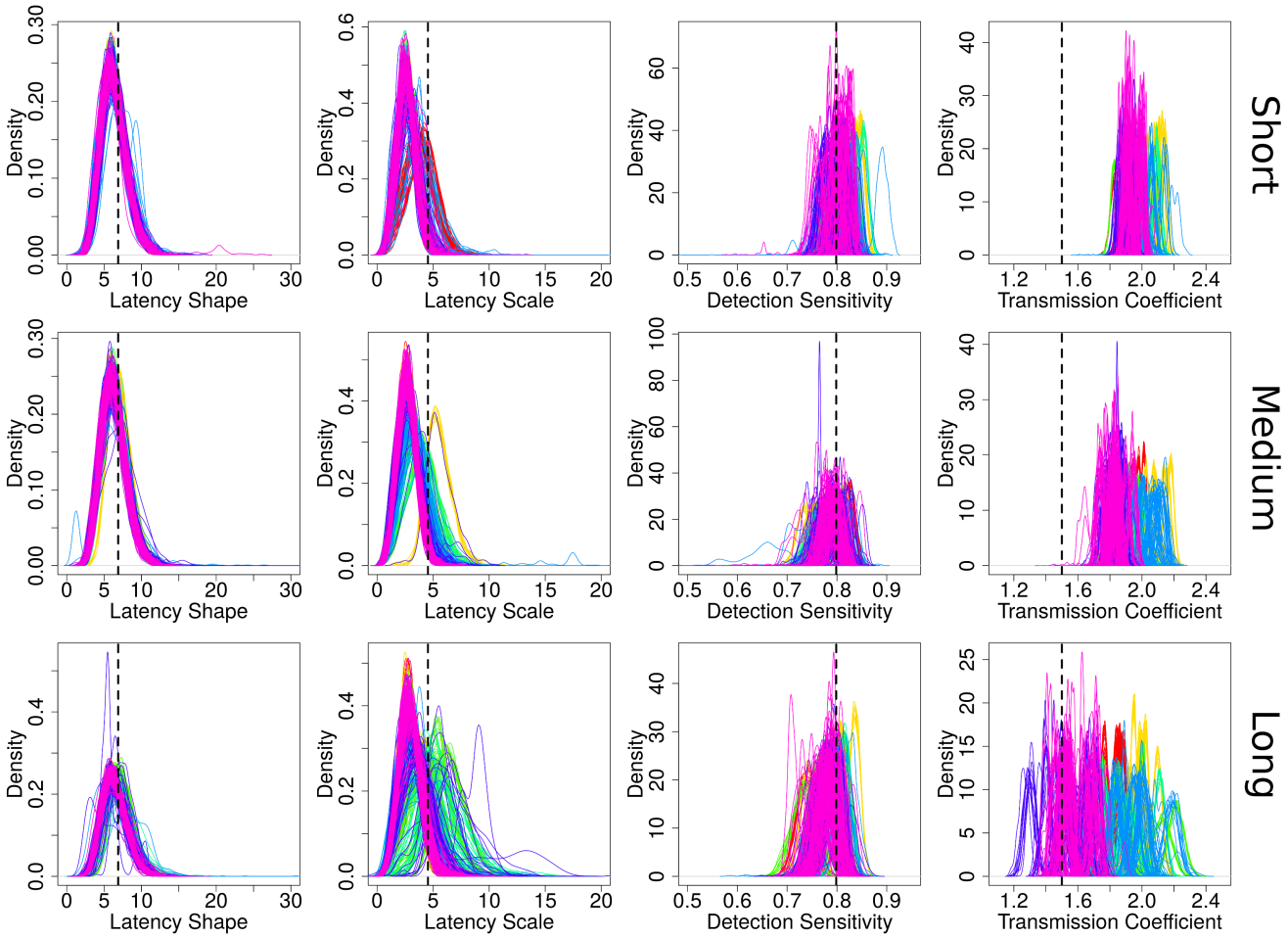

Fig S12. Comparison of simulated and estimated nuisance parameters. For each combination of short-, medium- and long-range kernels (from top to bottom) and introduction scenarios (colour-coded as in S3, S8, S9 and S10 Figs), 10 epidemics were simulated and 10 MCMC chains were run per simulated epidemic. The curves represent the posterior distribution of the parameters obtained under the most exhaustive estimation scheme (Θ_4) for all chains with non-negligible mean posterior likelihood. Dashed lines: parameter values used in the simulations.

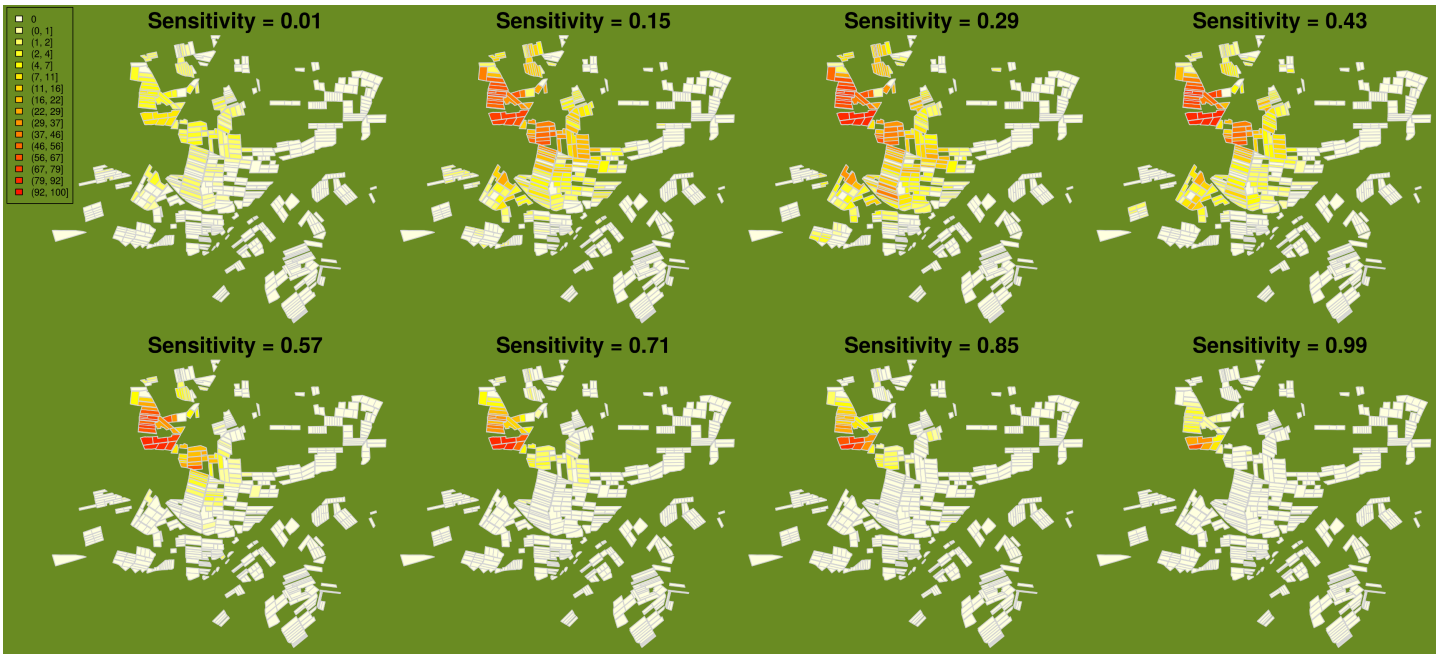

Fig S13. Cumulative detected incidence at the end of year 22 across the range of detection sensitivities (ρ) tested in the dedicated simulation study. Each polygon represents one peach orchard. All eight simulations start at year 1 from a unique introduction patch with 25% initial prevalence and spread is determined by the long-range kernel. Note that the final detected prevalence varies non-monotonically with detection sensitivity because the removal of detected trees reduces disease spread.

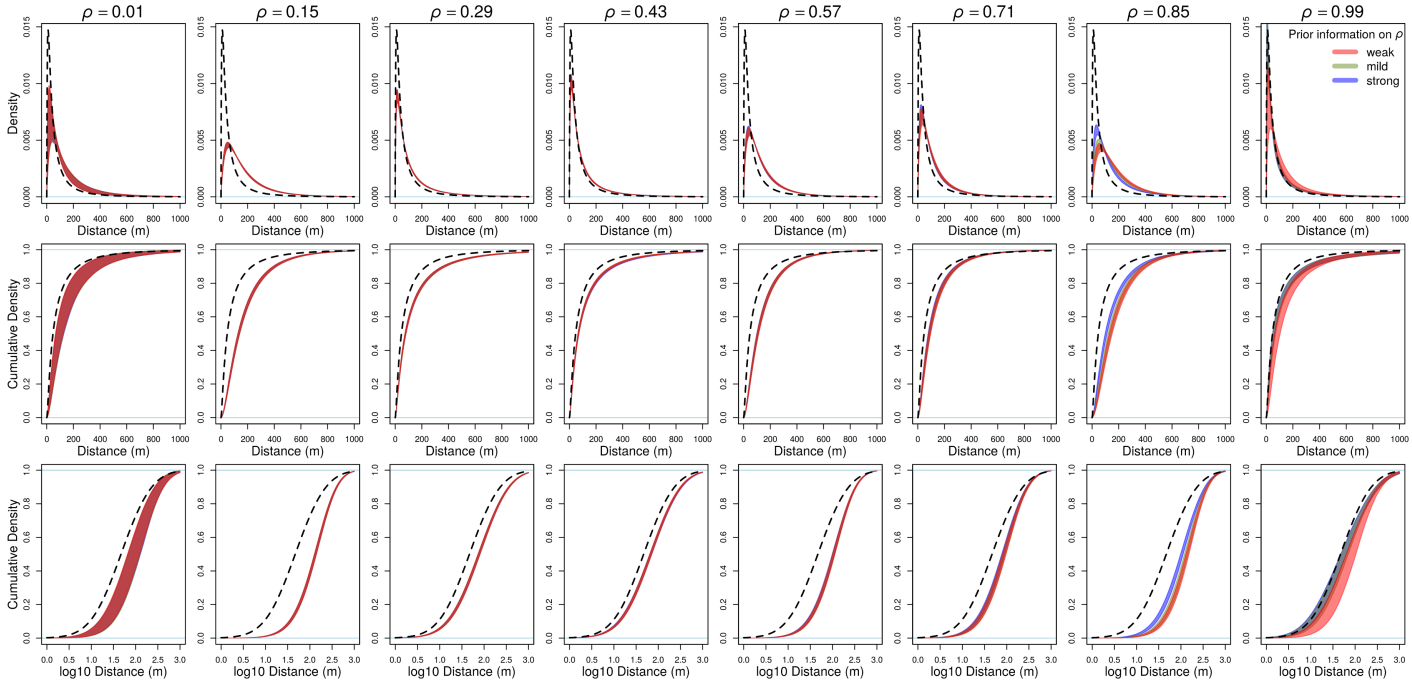

Fig S14. Influence of detection sensitivity on the estimation of the long-range dispersal kernel. For each detection sensitivity, a single epidemic was simulated using the long-range kernel (black dashed curve). The posterior distributions of the estimated kernels (obtained from all MCMC chains with non-negligible mean posterior likelihood) are shown for three levels of prior information. Kernels are represented by their marginal probability density function f^{1D} (top row), and by their marginal cumulative distribution function F^{1D} with the distance from the source represented on the natural scale (middle row) or on the $\log_{10}$ scale (bottom row).

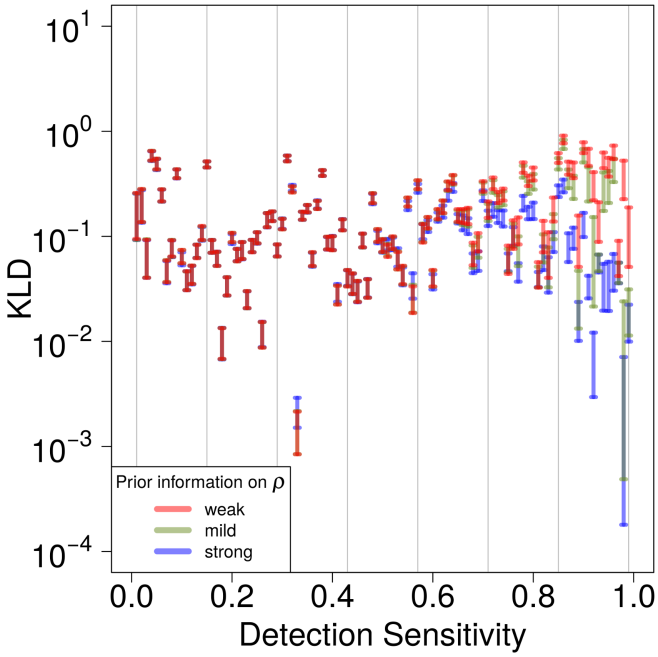

Fig S15. Influence of detection sensitivity on the distance between simulated and estimated long-range dispersal kernels. For each of the 99 detection sensitivities, a single epidemic was simulated using the long-range kernel. For three levels of prior information, each bar represents a 95% credibility interval on the Kullback-Leibler distance (KLD) between simulated and estimated dispersal kernels (obtained from all MCMC chains with non-negligible mean posterior likelihood). The grey vertical lines correspond to the values of detection sensitivity used in S13 and S14 Figs.

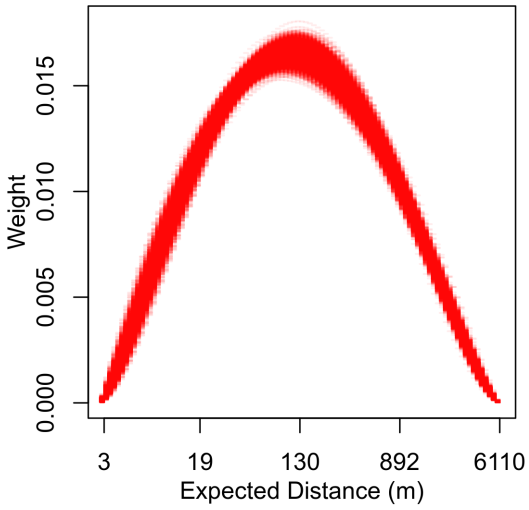

Fig S16. Estimated weights of the (BWME) dispersal kernel for the sharka epidemic. The posterior distribution of the weights (calculated with (Eq 11) for a mixture of 100 exponential kernels) is obtained for κ=11 (i.e. the number of introduction patches maximising the Fisher information). The plotted posterior distribution of weights (as a function of the expected distance of each kernel) was obtained from 4000 MCMC samples. One line is plotted per sample.

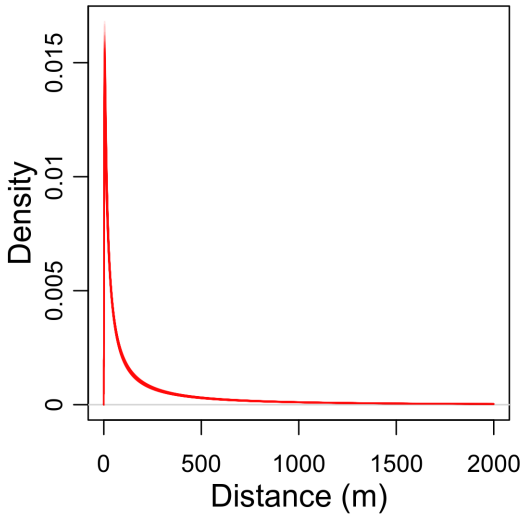

Fig S17. Estimated dispersal density for the sharka epidemic. The posterior distribution of the marginal probability density function, f^{1D}, of the fitted dispersal kernel, obtained for κ=11 (i.e. the number of introduction patches maximising the Fisher information). The plotted posterior distributions were obtained from 4000 MCMC samples. One line is plotted per sample.