Translating surveillance data into incidence estimates

Y. Bourhis[1], T. Gottwald[2], F. van den Bosch[1,3].

February 8, 2019

[1] Rothamsted Research, Department of Biointeraction and Crop Protection, Harpenden, AL5 2JQ, UK

[2] US Department of Agriculture, Agricultural Research Service, Ft. Pierce, Florida 34945, USA

[3] Department of Environment & Agriculture, Centre for Crop and Disease Management, Curtin University, Bentley 6102, Perth, Australia

for the *Philosophical Transactions of the Royal Society B*

Abstract

Monitoring a population for a disease requires the hosts to be sampled and tested for the pathogen. This results in sampling series from which we may estimate the disease incidence, *i.e.* the proportion of hosts infected. Existing estimation methods assume that disease incidence does not change between monitoring rounds, resulting in an underestimation of the disease incidence. In this paper we develop an incidence estimation model accounting for epidemic growth with monitoring rounds that sample varying incidence. We also show how to accommodate the asymptomatic period that is characteristic of most diseases. For practical use, we produce an approximation of the model, which is subsequently shown to be accurate for relevant epidemic and sampling parameters. Both the approximation and the full model are applied to stochastic spatial simulations of epidemics. The results prove their consistency for a very wide range of situations. The estimation model is made available as an online application.

Keywords: Disease Surveillance, Sampling Theory, Spatial Epidemiology

1 Introduction

Monitoring programs are used to keep track of the invasion and spread of human, animal and plant pathogens. They are often structured in discrete rounds of inspection, during which subsamples of the host population are assessed for disease status (Parnell et al., 2017). Given a sequence of monitoring rounds, a key question in interpreting these data is the estimation of the incidence[1] of the disease in the host population. There are two special cases of this general question that have received some attention.

Firstly, monitoring is often motivated by the need for early responses to enable eradication or containment. For example, *early detection* of the disease permits reduced culling of animal and plant hosts (Carpenter et al., 2011; Cunniffe et al., 2015, 2016), as well as limited deployments of emergency quarantines or travel restrictions for human hosts (applied *e.g.* for SARS, Smith, 2006). Secondly, monitoring is frequently motivated by the desire to *prove disease absence* from a host population (Caporale et al., 2012), which is important for the transport and trade of hosts. The main question then concerns the sufficient sample size (Cannon, 2002). An example of this is the practical "rule of three" (Louis, 1981; Hanley and Lippman-Hand, 1983). It gives the upper bound of the 95% confidence interval (CI) of the incidence when all of the N sampled hosts are assessed as healthy: $Q_{95} = 3/(N+1)$. Estimating disease incidence (noted q hereafter), or proving its absence, is mostly interesting during the early stages of epidemics, *i.e.* when incidences are low and containment measures are still promising.

Simple practices like the "rule of three" make the assumption that the samples are independent binomial draws with probability q and size N. However, as a pathogen spreads across the surveyed population, our samples will carry dependencies to the underlying epidemic process. For example, by pooling all the samples together, we neglect the fact that early monitoring rounds have most likely sampled a lower incidence q than the current one, resulting in an underestimation of the incidence. An alternative and unbiased solution is to estimate q only from the last round to date. But obviously, such a poor use of data would only be tolerable in cases where the monitoring interval and epidemic growth rate are both very large, so that the previous monitoring rounds can be deemed uninformative. The temporal dependence of samples has been addressed by Metz et al. (1983) in the design of appropriate monitoring programs, as well as by Parnell et al. (2015) and Bourhis et al. (2018) for the estimation of the disease incidence after the disease's *first discovery* or before its discovery (*disease absence*), respectively.

We propose here a generalised solution to the incidence estimation problem. Making use of all monitoring

[1] We use here the plant pathology definition where incidence is the fraction of host units infected. In human and other animal pathology this is termed prevalence.

data, it applies to the previously addressed *first dicovery* and *disease absence* cases, but extends to any monitoring outcome. Building on the simple logistic equation, our estimation model accounts for the progression of the disease during the monitoring period. Following the idea of the rule of three, and similar to Parnell et al. (2012) and Alonso Chavez et al. (2016), we also produce an approximation of this model. Its derivation only requires simple algebraic operations which makes it more suitable for practitioners than the full estimation model. Both the estimation model and its approximation are then confronted with epidemic simulations: first with non-spatial and deterministic simulations, and secondly with spatially explicit and stochastic simulations of epidemics running on contrasted distributions of hosts. The results support the accuracy and practical usefulness of the estimation model, which has subsequently been made available as an online software application.

2 Material and Methods

Monitoring a population for a disease results in sampling series as shown in Table 1. We define K as the number of monitoring rounds iterated in time. N_k is the sampling size of monitoring round k, *i.e.* the number of hosts whose pathological status is assessed at time t_k. M_k is the number infected hosts detected during round k. Finally, Δ_k is the time interval between monitoring rounds k and $k+1$.

Table 1: Variables and structure of a sampling series.

Monitoring round	1	2	...	k	...	$K-1$	K
Number of samples	N_1	N_2	...	N_k	...	N_{K-1}	N_K
Number of positives	M_1	M_2	...	M_k	...	M_{K-1}	M_K
Time interval	Δ_1	Δ_2	...	Δ_k	...	Δ_{K-1}	—

One monitoring round

Considering q the disease incidence in the population, the probability of M positive observations out of a sample of size N, is given by the binomial probability distribution

$$P(M|q;N) = \binom{N}{M}(1-q)^{N-M}q^M. \tag{1}$$

A more general form, accounting for the occurrences of false positives and negatives in the detection process, is

$$P(M|q;N) = \binom{N}{M}\left[(1-q)(1-\theta_{fp}) + q\theta_{fn})\right]^{N-M}\left[(1-q)\theta_{fp} + q(1-\theta_{fn})\right]^M, \tag{2}$$

where θ_{fn} and θ_{fp} are respectively the rates of false negatives and false positives (Cameron and Baldock, 1998a). For simplicity, the following developments do not explicitly incorporate those rates, which are nonetheless part of the estimation model provided in the application.

In a practical context, q is the variable that we want to estimate from samples characterised by their size N and their outcome M. To this end we use Bayes' rule:

$$P(q|M;N) = \frac{P(q)P(M|q;N)}{\int_0^1 P(q)P(M|q;N)dq}, \tag{3}$$

where $P(q|M,N)$ is the probability distribution of q given M and N. Assuming no information on the incidence before sampling, we set a uniform prior $P(q)$, resulting in $P(q|M;N) \propto P(M|q;N)$ (Gelman et al., 2003).

K monitoring rounds

To account properly for the dynamic incidence between monitoring rounds, we inform the binomial probability distribution with an epidemiological component Z_k (as in Hamelin et al., 2016, whose maximum-likelihood approach is equivalent to our Bayesian formulation with flat priors). Z_k gives the relation between q_K, the incidence to estimate, and q_k, the incidence at sampling time t_k, as $q_k = Z_k q_K$. Hence,

$$P(\boldsymbol{M}|q_K;\boldsymbol{N}) = \prod_{k=1}^{K} \binom{N_k}{M_k} (1 - Z_k q_K)^{N_k - M_k} (Z_k q_K)^{M_k}, \tag{4}$$

where $\boldsymbol{M}$ and $\boldsymbol{N}$ on the left-hand side represents the whole sampling series, *i.e* M_1, M_2,...,M_K and N_1, N_2,...,N_K.

We assume that the disease incidence, q, evolves logistically (van der Plank, 1963; Murray, 2002) between times t_k and t_K:

$$q_K = \frac{q_k e^{r(t_K - t_k)}}{1 + q_k(e^{r(t_K - t_k)} - 1)}, \tag{5}$$

where r is the epidemic growth rate, while q_k and q_K respectively stand for $q(t_k)$ and $q(t_K)$. For simplicity, we hereafter express time relative to $t_K = 0$, the time of both the last sampling round and the estimation. Hence, we define Z_k as:

$$Z_k = \frac{q(t_k)}{q(t_K)} = \frac{q_K e^{r t_k}}{1 + q_K(e^{r t_k} - 1)} \Big/ q_K = \frac{e^{r t_k}}{1 + q_K(e^{r t_k} - 1)}, \tag{6}$$

where $t_k < 0$ as $t_K = 0$.

Similarly to the case of one monitoring round, we use Bayes' rule to get the unnormalised posterior distribution $P(q_K|\boldsymbol{M};\boldsymbol{N})$. Practically, it is given by Eq. 4, which is computed for a discretised array of $q \in [0,1]$, and from which quantiles Q_X can be derived (a method called grid approximation, see *e.g.* Kruschke, 2014).

A useful approximation

The upper bound of the CI is a useful measure of the highest, still likely, incidence we can expect in the population given the outcome of a monitoring program. We propose in this section, an approximation of this quantity not requiring the derivation of the full probability distribution of $P(q_K|\boldsymbol{M};\boldsymbol{N})$. Various methods exist for approximating the CI of a binomial parameter (Wallis, 2013). After preliminary testing of those methods against the binomial-shaped probability density given by Eq. 4, we choose the Agresti-Coull interval for its accuracy for low incidences (Agresti and Coull, 1998). Therefore, the approximated upper limit of the $X\%$ CI is defined as

$$\tilde{Q}_X = \min\left(1,\ \tilde{p} + z\sqrt{\max\left(0, \frac{\tilde{p}}{N+z^2}\left(1-\tilde{p}\right)\right)}\right), \tag{7}$$

where

$$\tilde{p} = \frac{1}{N+z^2}\left(M + \frac{z^2}{2}\right), \tag{8}$$

and where z is the corresponding $1-\alpha/2$ quantile of the standard normal distribution (with α the probability of type I error). For the one-sided 95% CI, we derive $\tilde{Q}_{95}$ by setting $z = 1.645$.

The approximated $\tilde{Q}_X$ also needs to account for the epidemic growth. As previously with Z_k, we now define $\tilde{Z}_k$ to quantify the disease evolution between rounds. In this case, we are unable to use the logistic model because its non-linearity makes the derivation of $\tilde{Q}_X$ intractable. This, however, was no concern for the full model and the grid approximation method used to derive $P(q_K|\boldsymbol{M};\boldsymbol{N})$. Consequently, approximating the logistic growth model by its exponential variant, $\tilde{Z}_k$ is given by

$$\tilde{Z}_k = e^{rt_k} = \exp\left(-r\sum_{i=k}^{K}\Delta_i\right). \tag{9}$$

In practice, the exponential assumption is realistic as, during early infection, the epidemic growth is exponential, even according to the logistic model (van der Plank, 1963). Finally, we aggregate the samples together with respect to the epidemic growth via $\tilde{Z}_k$:

$$M = \sum_{k=1}^{K} M_k, \quad \text{and} \quad N = \sum_{k=1}^{K} N_k\tilde{Z}_k. \tag{10}$$

These aggregated values of M and N are then substituted in Eqs. 8 and 7 to derive $\tilde{Q}_X$. Scaling only N_k with Z_k has two effects: (1) the historic sampling rounds k contribute less than the recent ones to the reduction of the uncertainty (reduced sample size N); and (2) the sampling rounds k that include detection events ($M_k > 0$) see their contribution to $\tilde{Q}_X$ increased (larger M/N), hence accounting for the putative spread of the disease from those M_k infected hosts between times t_k and t_K. The *min* and *max* operators in Eq. 7 are added to deal with the possibility of having $N < M$ for some values of $\tilde{Z}_k$.

As mentioned in the introduction, this estimation model and its approximation cover the two specific contexts of *first discovery* and *disease absence* addressed respectively by Parnell et al. (2012) and Bourhis et al. (2018) (see Supplementary Materials for details). Their strength is however, that they extend to any sampling series, no matter its outcome M_k and regularity in sampling size or frequency.

Asymptomatic period

For most diseases, infected hosts develop symptoms after an asymptomatic (or incubation) period (Thompson et al., 2016). Often, asymptomatic hosts contribute to the epidemic dynamics by spreading the disease while still undetectable (cryptic) when sampled. The logistic equation handles this period, noted σ, as in Alonso Chavez et al. (2016):

$$q_T(t_K) = \frac{q(t_k)e^{r(t_K - t_k + \sigma)}}{1 + q(t_k)(e^{r(t_K - t_k + \sigma)} - 1)}. \tag{11}$$

This relates the total incidence q_T at the last sampling round t_K (*i.e.* the quantity to estimate) to the detectable incidences at the different sampling times $q(t_k)$ (*i.e.* the sampled quantities). Hence, Z_k becomes:

$$Z_k = \frac{q(t_k)}{q_T(t_K)} = \frac{e^{r(t_k - \sigma)}}{1 + q_T(t_K)(e^{r(t_k - \sigma)} - 1)}. \tag{12}$$

For the exponential approximation, Eq. 9 simply becomes

$$\tilde{Z}_k = e^{r(t_k - \sigma)}. \tag{13}$$

Testing the model

The consistency of the full model and the accuracy of its approximation are first tested against simulations of stochastic sampling on non-spatial logistic epidemics. We consider a uniform distribution of incidences q_T that we want to estimate individually. For each one of them, an epidemic is simulated until incidence q_T is reached and a monitoring program is designed with N_k and Δ_k drawn from Poisson distributions of mean $\overline{N}$ and $\overline{\Delta}$ respectfully. From the logistic equation (Eq. 5), the detectable incidence q is derived for every sampling date t_k. Then binomial draws with probability $p = q(t_k)$ and size $n = N_k$ simulate the sampling process of the hosts, resulting in M_k. For every q_T an exact upper bound of its CI, Q_X, is derived with the full model, while an approximated one, $\tilde{Q}_X$, is derived with the approximation. To test our model we check that the upper limits of the $X\%$ CI are above q_T in $X\%$ of cases. This test is done for contrasted values of the sampling ($\overline{N}$ and $\overline{\Delta}$) and epidemic parameters (r and σ).

The full model and its approximation are also tested against spatially explicit and stochastic epidemic simulations (as in Hyatt-Twynam et al., 2017). In this case, the epidemics are no longer modelled with the

logistic equation but through a transmission rate and a dispersal kernel of the pathogens. To this end, the hosts are distributed in a 2D-space and aggregated randomly in field-like structures mimicking the distribution of the trees in an orchard. Details of this landscape model are given as Supplementary Material. Transmission is governed by an exponential power kernel (Rieux et al., 2014). The probability of a susceptible individual becoming infected in a unit of time is then given by

$$p(s \in S) = \beta \frac{b\mathcal{A}}{2\pi\theta^2\Gamma(2/b)} \sum_{i \in I} \exp(-|\boldsymbol{x}_i - \boldsymbol{x}_s|^b/\theta^b), \tag{14}$$

where s is a susceptible host among the set of all susceptible hosts S. Similarly, i and I represent the infected hosts. $\mathcal{A}$ is the area occupied by one host and Γ is the Gamma function. β is the probability of infection, θ is the dispersal scale and b is a shape parameter (producing fat-tailed kernels for $b < 1$). The coordinates $\boldsymbol{x}$ mark the locations of the hosts. Following Klein et al. (2006), the mean dispersal distance for this 2D kernel is given by:

$$\delta = \theta\ \Gamma(3/b)\ /\ \Gamma(2/b). \tag{15}$$

These epidemics are simulated with the τ-leap version of the Gillespie stochastic simulation algorithm (see *e.g.* Keeling and Rohani, 2008). The estimation model and its approximation are evaluated in the same way as the non-spatial case.

3 Results

Model behaviours

Figure 1 illustrates the effects of the epidemic and sampling parameters on the resulting probability distributions of the incidence and upper quantiles Q_{95}. Increasing M_k, the number of detected infected hosts in the sample, unsurprisingly increases the estimated incidence. Increasing the sample size N_k reduces the uncertainty in the estimates. Increasing the sampling interval Δ decreases the impact of the historic samples on the estimation. This reflects the fact that samples taken further back in time are less informative of current disease incidence. As for the epidemic parameters, the growth rate r and the asymptomatic period σ (not shown on Figure 1 for dimensional reasons) have very similar effects to Δ. Increasing them increases the estimated incidence by decreasing the impact of the historic samples (which are the ones sampling lower incidences q). Increasing any of the parameters Δ, r or σ also reduces the effective sample size (*i.e.* $\sum_{k=1}^{K} N_k Z_k$), which increases the uncertainty on the estimates (*i.e.* producing probability distributions with larger variance).

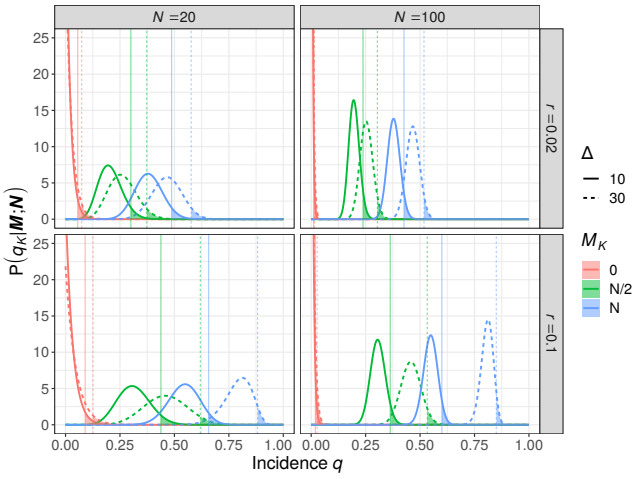

Figure 1: Probability distributions of the incidence q given by Eqs. 4 and 3. The vertical lines mark Q_{95}, the upper limit of the 95% CI. The distributions result from a sampling series composed of $K = 3$ monitoring rounds, of which the first two are fully negative (*i.e.* $M_1 = M_2 = 0$) and the last varies from $M_3 = 0$ (*i.e.* all sampled hosts are negative) to $M_3 = N$ (*i.e.* all sampled hosts are positive). These probability distributions are represented for varying values of epidemic growth rate r, sampling size N and sampling interval Δ.

Test against logistic epidemics

Figure 2 shows the distribution of the exact and approximated upper bounds of the 95% CI, Q_{95} and $\tilde{Q}_{95}$, for uniform distributions of q_T and different values of the epidemic and sampling parameters. The full model, which similarly to the simulations builds on the logistic equation, behaves exactly as expected: it ensures that 95% of the Q_{95} are above their respective q_T, for every set of parameters tested. On the other hand, the approximation displays another behaviour which is explained by its underlying exponential growth model. For the low incidences which are relevant to practice (*i.e.* say $q_T < 0.25$), the approximation is accurate (the distributions of Q_{95} and $\tilde{Q}_{95}$ do overlap). For higher incidences, *i.e.* when the logistic growth decelerates unlike the exponential growth, the approximation overestimates the incidence (increasingly with r, σ and Δ).

Another model behaviour of particular interest occurs when r and σ are large (see the rightmost column of Figure 2). We observe that the estimated Q_{95} and $\tilde{Q}_{95}$ do not align well with the diagonal for small incidences q_T. For those cases of very hazardous pathogens with high epidemic growth rates and long asymptomatic periods, the sampling size N is too small to allow discrimination between the non-detection cases (*i.e.* the one for which all the $M_k = 0$), and a larger sampling effort is needed for the estimation to be informative.

Although increasing r and σ accelerates the divergence between the logistic and the exponential curves, the approximation appears accurate for early infections even considering very high values of epidemic parameters such as $r = 0.1$ day^{-1} or $\sigma = 100$ days.

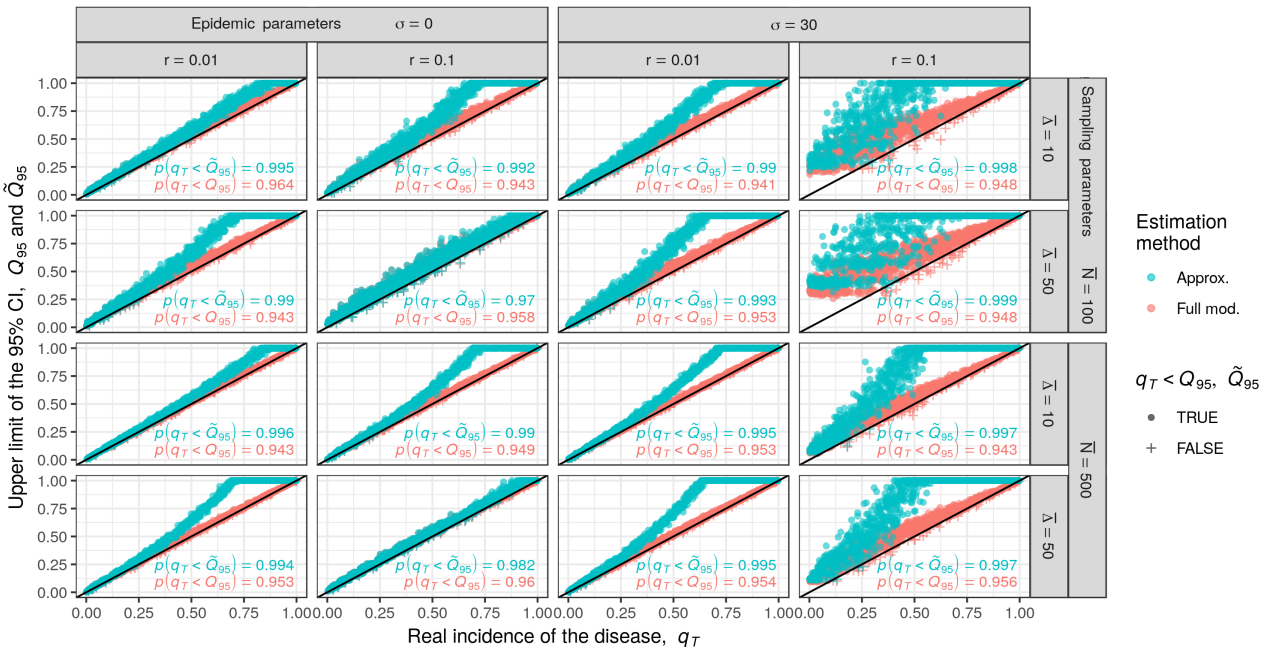

Figure 2: Estimation of Q_{95} and $\tilde{Q}_{95}$ from sampling series of non-spatial epidemics, *i.e.* simulated with the logistic equation (Eq. 5). These estimations are made for contrasted values of sampling and epidemic parameters (and for $K = 5$ monitoring rounds). Using here the 95% CI, we expect 95% of the estimated Q_{95} and $\tilde{Q}_{95}$ to be above the actual incidence in the field at the end of monitoring q_T, *i.e.* above the oblique black line. The inserted texts summarise these scores for the full model (in red) and its approximation (in blue).

Test against spatial epidemics

When locating the hosts in space, the epidemic becomes driven by two new elements: the dispersal range of the pathogen and the intensity of host clustering (Brown and Bolker, 2004). Both determine how easily the pathogen spreads across the landscape or remains restricted to a local group of hosts. Random distributions of hosts and long dispersal ranges result in smooth progressions of the pathogen across the landscape, following a logistic-like curve. However, as the dispersal range decreases and host aggregation increases, the simulated epidemics will tend to include interruptions between periods of seemingly logistic growth within host clusters. Questions then arise regarding the performance of our estimation model on such epidemics.

The estimation model and its approximation are tested for varying host aggregations and dispersal ranges. Host aggregation is summarised by μ, the number of hosts in a field (*sensu* host cluster). For a given landscape-scale population of hosts, more hosts per field means fewer but more populated fields (see the Supplementary Material for an illustration). The dispersal scale θ is translated in terms of mean dispersal distance δ (see Eq. 15), while μ is translated in terms of $\overline{d}$, a landscape metric measuring the mean minimal distance between the fields within a landscape (see Euclidean Nearest Distance in Leitao et al., 2006).

Similar to Figure 2, Figure 3 shows the performance of the model and its approximation for gradients of

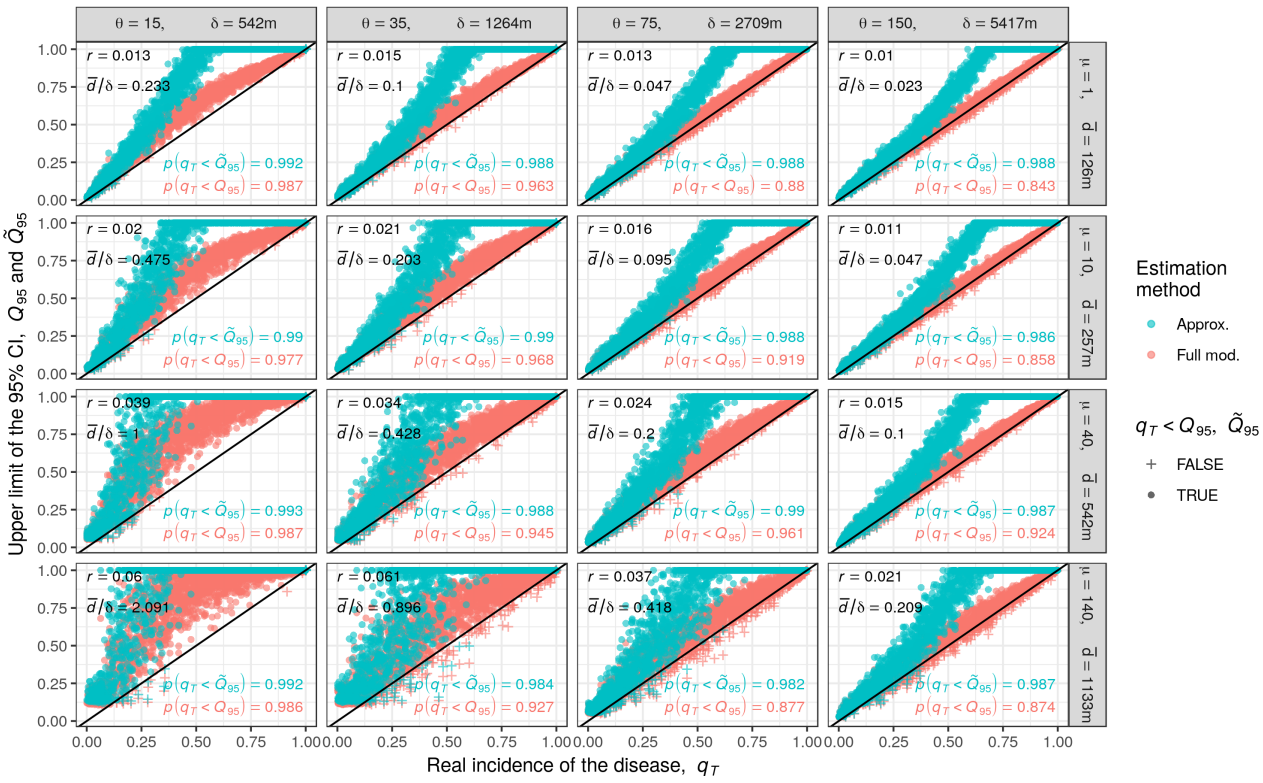

Figure 3: Estimation of Q_{95} and $\tilde{Q}_{95}$ from sampling series realised on spatially explicit epidemics, *i.e.* simulated with the dispersal kernel (Eq. 14). These estimations are made for varying dispersal ranges θ and hosts aggregations μ, while maintaining constant values of the non-spatial parameters ($N = 100$, $\Delta = 30$, $\sigma = 30$, $K = 5$, as well as $\beta = 75$ and $b = 0.45$ for the remaining kernel parameters). For better understanding, θ and μ are shown with their distance translation in meters, δ and $\overline{d}$. The identified logistic growth rate r is given for each experiment. The resulting distributions of Q_{95} and $\tilde{Q}_{95}$ are qualitatively similar for other realistic values of the fixed parameters.

dispersal scales θ (columns) and host aggregations μ (rows). For each parameter set θ and μ (*i.e.* each panel in Figure 3), 50 epidemics are first simulated for 50 different landscapes in order to identify the value of r that produces the best fitting logistic curve (with least squares). This r then informs the incidence estimation model and its approximation for the subsequent testing set of 2000 epidemics and landscapes. Most of Figure 3 agrees with expectations: the estimated Q_{95} align neatly above the diagonal, showing in practice the accuracy of the estimation model. The approximation appears to be a good simplification of the full model for early detection. However, the estimation model also produces overestimations of the incidence, specifically in the bottom row and left column (*i.e.* where the dots do not align above the diagonal). These are cases of epidemics for which the distance between host clusters (quantified by $\overline{d}$) is too large for the pathogen dispersal range (quantified by δ), hence producing unsteady progressions of the pathogens across the landscapes. This illustrates that our model is of limited interest in such cases where $\overline{d}/\delta \leq 0.5$.

We notice also that $p(q_T < Q_{95})$ can be below the 95% expectation. This results from the fact that the

stochasticity of the simulations scatters the realised epidemic curves symmetrically around the fitted logistic one (whose identified parameter r is subsequently used by the estimation model). This is no concern in practice where the epidemic parameters are taken conservatively from previous observations of similar outbreaks (*e.g* highest observed values of r or σ). Here we choose central estimates (through least squares) for illustrative purposes. Nonetheless, parameter uncertainty can be accounted for in the online application assuming that r and σ can be described with normal distributions. On how to deal with epidemic parameter uncertainty, see Neri et al. (2014); Hyatt-Twynam et al. (2017).

4 Discussion

The model developed in this paper is suitable for many monitoring designs, including those with irregular sampling sizes and time intervals between rounds. The model weights each monitoring outcome according to an estimate of the population incidence at their respective sampling time, before aggregating them into a single binomial-shaped probability distribution of the incidence. The quantiles of this distribution have practical interests for policy-makers. The model is directly applicable for situations in which surveillance does not depend on the self-reporting of symptomatic hosts, which makes it appropriate for most animal and plant species. Our model is also appropriate for certain monitoring schemes aimed at pathogens of humans, for example visitations of rural villages to find Ebola infections where access to healthcare is limited (Namukose et al., 2018; Thompson et al., 2019).

Calculating the probability density of the incidence from the sampling series is computationally inexpensive, but still requires technical proficiency. Therefore, we have produced an online application interfacing the full model as exhaustively as possible, as well as an approximation of the model which can be derived with simple algebraic operations. Our intention is to equip the widest audience of practitioners with this incidence estimation capability. The approximation is as flexible as the original model, and we have shown that its inaccuracies are restricted to high level of incidences that are less relevant when dealing with emerging epidemics. However, in case such high incidence estimation is needed, we have seen that the approximation is conservative, *i.e.* biased towards an overestimation of disease progress (which is not always acceptable, since it might lead to overzealous control, see *e.g.* Thompson et al., 2018).

The model relies on the simple and deterministic logistic equation. That it is consistent with more complex systems is not obvious. The tests presented here against spatial and stochastic simulations of epidemics show that our non-spatial model is robust to the significant deviations from the logistic equation, products of both

spatiality and stochasticity. The model gives accurate estimates of the disease incidence for most simulated epidemics considered here. However, for highly aggregated host distributions and short distance dispersing pathogens, the deviation from the logistic equation can be too great. In those contexts, the disease progression across the landscape is not steady but punctuated by rare events: the pathogen jumps between distant host clusters. Then, the very distinctive trajectories this epidemic can take do not simplify well into a single logistic curve. In such cases, reduced pathogen dispersal and increased host aggregation result in habitat fragmentation for the pathogen. The estimation should then be attempted on individual clusters or a multiscale approach considered (as in Cameron and Baldock, 1998b; Coulston et al., 2008).

From plants to animals, the major shift regarding epidemics lies in individual movement. In many cases, this can be overlooked as it does not necessarily imply movement of the sampling units (*e.g.* herds/farms in Bates et al., 2003). When sampling individuals however, our model is applicable to well-mixed populations, *i.e.* where the pathogen spread is steady and not too impacted by spatio-temporal structure in the host population. We saw the limits of this assumption in Figure 3 where highly clustered distributions of hosts cause significant deviations from model predictions. Such deviations may, for example, be increased if clustering correlates with heterogeneous susceptibility of hosts (*e.g.* age-related aggregation like schools), or attenuated by mutable clusters (*e.g.* commuting).

Recent technological innovations are changing epidemiological surveillance for more timely and exhaustive censuses. For example, the monitoring of human epidemics is already augmented by the supervision of social networks (Chen et al., 2014) and internet search queries (Yuan et al., 2013; Yang et al., 2015). Tree monitoring could also be assisted by satellite high-resolution imagery (Li et al., 2014; Salgadoe et al., 2018). Those innovations will still need robust and epidemiologically informed estimation methods and, even if monitoring is conducted continuously, there is no reason to see them incompatible with an adaptation of our model. However, in any foreseeable future, most contagions will still be monitored through discrete and censored inspections and hence, remain within the immediate scope of the estimation model presented here.

Acknowledgements

The work at Rothamsted forms part of the Smart Crop Protection (SCP) strategic programme (BBS/OS/CP/000001) funded through the Biotechnology and Biological Sciences Research Council's Industrial Strategy Challenge Fund. Authors are also thankful to the US Department of Agriculture for funding support. We are grateful to Francisco Lopez-Ruiz for the idea to make the model available as an application. Finally, we are grateful to

Robin Thompson and three anonymous reviewers whose time and efforts greatly improved the manuscript.

Supplementary Materials

A The estimation model as an online application.

B Details, illustration and code for the landscape model.

C Development of the specific approximations for first discovery and disease absence.

References

Agresti, A. and Coull, B. A. (1998). Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126.

Alonso Chavez, V., Parnell, S., and van den, F. (2016). Monitoring invasive pathogens in plant nurseries for early-detection and to minimise the probability of escape. *Journal of Theoretical Biology*, 407:290–302.

Bates, T. W., Thurmond, M. C., and Carpenter, T. E. (2003). Description of an epidemic simulation model for use in evaluating strategies to control an outbreak of foot-and-mouth disease. *American Journal of Veterinary Research*, 64(2):195–204.

Bourhis, Y., Gottwald, T. R., Lopez-Ruiz, F. J., Patarapuwadol, S., and van den Bosch, F. (2018). Sampling for disease absence—deriving informed monitoring from epidemic traits. *Journal of Theoretical Biology*.

Brown, D. H. and Bolker, B. M. (2004). The effects of disease dispersal and host clustering on the epidemic threshold in plants. *Bulletin of Mathematical Biology*, 66(2):341–371.

Cameron, A. R. and Baldock, F. C. (1998a). A new probability formula for surveys to substantiate freedom from disease. *Preventive Veterinary Medicine*, 34(1):1–17.

Cameron, A. R. and Baldock, F. C. (1998b). Two-stage sampling in surveys to substantiate freedom from disease. *Preventive Veterinary Medicine*, 34(1):19–30.

Cannon, R. M. (2002). Demonstrating disease freedom—combining confidence levels. *Preventive Veterinary Medicine*, 52(3):227–249.

Caporale, V., Giovannini, A., and Zepeda, C. (2012). Surveillance strategies for foot and mouth disease to prove absence of disease and absence of viral circulation: -EN- -FR- Les stratégies de surveillance de la fièvre aphteuse visant à démontrer l'absence de la maladie et l'absence de circulation virale -ES- Estrategias de vigilancia de la fiebre aftosa para demostrar la ausencia de enfermedad y de circulación de virus. *Revue Scientifique et Technique de l'OIE*, 31(3):747–759.

Carpenter, T. E., O'Brien, J. M., Hagerman, A. D., and McCarl, B. A. (2011). Epidemic and Economic Impacts of Delayed Detection of Foot-And-Mouth Disease: A Case Study of a Simulated Outbreak in California. *Journal of Veterinary Diagnostic Investigation*, 23(1):26–33.

Chen, L., Hossain, K. T., Butler, P., Ramakrishnan, N., and Prakash, B. A. (2014). Flu Gone Viral: Syndromic Surveillance of Flu on Twitter Using Temporal Topic Models. In *2014 IEEE International Conference on Data Mining*, pages 755–760. IEEE.

Coulston, J. W., Koch, F. H., Smith, W. D., and Sapio, F. J. (2008). Invasive forest pest surveillance: survey development and reliability. *Canadian Journal of Forest Research*, 38(9):2422–2433.

Cunniffe, N. J., Cobb, R. C., Meentemeyer, R. K., Rizzo, D. M., and Gilligan, C. A. (2016). Modeling when, where, and how to manage a forest epidemic, motivated by sudden oak death in California. *Proceedings of the National Academy of Sciences*, page 201602153.

Cunniffe, N. J., Stutt, R. O. J. H., DeSimone, R. E., Gottwald, T. R., and Gilligan, C. A. (2015). Optimising and Communicating Options for the Control of Invasive Plant Disease When There Is Epidemiological Uncertainty. *PLOS Computational Biology*, 11(4):e1004211.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis, Second Edition*. CRC Press. Google-Books-ID: TNYhnkXQSjAC.

Hamelin, F. M., Bisson, A., Desprez-Loustau, M.-L., Fabre, F., and Mailleret, L. (2016). Temporal niche differentiation of parasites sharing the same plant host: oak powdery mildew as a case study. *Ecosphere*, 7(11):e01517.

Hanley, J. A. and Lippman-Hand, A. (1983). If Nothing Goes Wrong, Is Everything All Right?: Interpreting Zero Numerators. *JAMA*, 249(13):1743–1745.

Hyatt-Twynam, S. R., Parnell, S., Stutt, R. O. J. H., Gottwald, T. R., Gilligan, C. A., and Cunniffe, N. J. (2017). Risk-based management of invading plant disease. *New Phytologist*, 214(3):1317–1329.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. Google-Books-ID: G8enmS23c6YC.

Klein, E. K., Lavigne, C., and Gouyon, P.-H. (2006). Mixing of propagules from discrete sources at long distance: comparing a dispersal tail to an exponential. *BMC Ecology*, 6:3.

Kruschke, J. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic Press. Google-Books-ID: FzvLAwAAQBAJ.

Leitao, A. B., Miller, J., Ahern, J., and McGarigal, K. (2006). *Measuring Landscapes: A Planner's Handbook*. Island Press, Washington, D.C.

Li, H., Lee, W. S., Wang, K., Ehsani, R., and Yang, C. (2014). 'Extended spectral angle mapping (ESAM)' for citrus greening disease detection using airborne hyperspectral imaging. *Precision Agriculture*, 15(2):162–183.

Louis, T. A. (1981). Confidence Intervals for a Binomial Parameter after Observing No Successes. *The American Statistician*, 35(3):154–154.

Metz, J. A. J., Wedel, M., and Angulo, A. F. (1983). Discovering an Epidemic before It Has Reached a Certain Level of Prevalence. *Biometrics*, 39(3):765–770.

Murray, J. D. (2002). *Mathematical Biology: I. An Introduction*. Interdisciplinary Applied Mathematics. Springer-Verlag, New York, 3 edition.

Namukose, E., Bowah, C., Cole, I., Dahn, G., Nyanzee, P., Saye, R., Duworko, M., Nsubuga, P., Mawanda, M., Mahmoud, N., Clement, P., Ngabirano, T. D., Nyenswah, T., and Gasasira, A. (2018). Active Case Finding for Improved Ebola Virus Disease Case Detection in Nimba County, Liberia, 2014/2015: Lessons Learned. *Advances in Public Health*, 2018:1–7.

Neri, F. M., Cook, A. R., Gibson, G. J., Gottwald, T. R., and Gilligan, C. A. (2014). Bayesian Analysis for Inference of an Emerging Epidemic: Citrus Canker in Urban Landscapes. *PLOS Computational Biology*, 10(4):e1003587.

Parnell, S., Gottwald, T., Gilks, W., and van den Bosch, F. (2012). Estimating the incidence of an epidemic when it is first discovered and the design of early detection monitoring. *Journal of Theoretical Biology*, 305:30–36.

Parnell, S., Gottwald, T. R., Cunniffe, N. J., Alonso Chavez, V., and van den Bosch, F. (2015). Early detection surveillance for an emerging plant pathogen: a rule of thumb to predict prevalence at first discovery. *Proceedings of the Royal Society B: Biological Sciences*, 282(1814):20151478.

Parnell, S., van den Bosch, F., Gottwald, T., and Gilligan, C. A. (2017). Surveillance to Inform Control of Emerging Plant Diseases: An Epidemiological Perspective. *Annual Review of Phytopathology*, 55(1):591–610.

Rieux, A., Soubeyrand, S., Bonnot, F., Klein, E. K., Ngando, J. E., Mehl, A., Ravigne, V., Carlier, J., and Bellaire, L. d. L. d. (2014). Long-Distance Wind-Dispersal of Spores in a Fungal Plant Pathogen: Estimation of Anisotropic Dispersal Kernels from an Extensive Field Experiment. *PLOS ONE*, 9(8):e103225.

Salgadoe, A., Robson, A., Lamb, D., Dann, E., and Searle, C. (2018). Quantifying the Severity of Phytophthora Root Rot Disease in Avocado Trees Using Image Analysis. *Remote Sensing*, 10(2):226.

Smith, R. D. (2006). Responding to global infectious disease outbreaks: Lessons from SARS on the role of risk perception, communication and management. *Social Science & Medicine*, 63(12):3113–3123.

Thompson, R., Morgan, O., and Jalava, K. (2019). Rigorous surveillance is necessary for high confidence in end-of-outbreak declarations for Ebola and other infectious diseases. *Philos Trans R Soc Lond B Biol Sci.*

Thompson, R. N., Gilligan, C. A., and Cunniffe, N. J. (2016). Detecting Presymptomatic Infection Is Necessary to Forecast Major Epidemics in the Earliest Stages of Infectious Disease Outbreaks. *PLOS Computational Biology*, 12(4):e1004836.

Thompson, R. N., Gilligan, C. A., and Cunniffe, N. J. (2018). Control fast or control smart: When should invading pathogens be controlled? *PLOS Computational Biology*, 14(2):e1006014.

van der Plank, J. E. (1963). *Plant Diseases: Epidemics and Control.* Academic Press, New York. Google-Books-ID: HqzSBAAAQBAJ.

Wallis, S. (2013). Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods. *Journal of Quantitative Linguistics*, 20(3):178–208.

Yang, S., Santillana, M., and Kou, S. C. (2015). Accurate estimation of influenza epidemics using Google search data via ARGO. *Proceedings of the National Academy of Sciences*, 112(47):14473–14478.

Yuan, Q., Nsoesie, E. O., Lv, B., Peng, G., Chunara, R., and Brownstein, J. S. (2013). Monitoring Influenza Epidemics in China with Search Query from Baidu. *PLOS ONE*, 8(5):e64323.