Early detection surveillance for an emerging plant pathogen: a rule of thumb to predict prevalence at first discovery

S. Parnell1,2, T. R. Gottwald3, N. J. Cunniffe4, V. Alonso Chavez1 and F. van den Bosch1

1Computational and Systems Biology, Rothamsted Research, Harpenden, Herts, AL5 2JQ, UK
2School of Environment and Life Sciences, University of Salford, Manchester, M5 4WT, UK
3USDA Agricultural Research Service, 2001 South Rock Road, Ft. Pierce, FL 34945, USA
4Department of Plant Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EA, UK

Emerging plant pathogens are a significant problem for conservation and food security. Surveillance is often instigated in an attempt to detect an invading epidemic before it gets out of control. Yet in practice many epidemics are not discovered until already at a high prevalence, partly due to a lack of quantitative understanding of how surveillance effort and the dynamics of an invading epidemic relate. We test a simple rule of thumb to determine, for a surveillance programme taking a fixed number of samples at regular intervals, the distribution of the prevalence an epidemic will have reached on first discovery (discovery-prevalence) and its expectation $E(q^*)$. We show that $E(q^*) = r/(N/D)$, i.e. simply the rate of epidemic growth divided by the rate of sampling; where $r$ is the epidemic growth rate, $N$ is the sample size and $D$ is the time between sampling rounds. We demonstrate the robustness of this rule of thumb using spatio-temporal epidemic models as well as data from real epidemics. Our work supports the view that, for the purposes of early detection surveillance, simple models can provide useful insights in apparently complex systems. The insight can inform decisions on surveillance resource allocation in plant health and has potential applicability to invasive species generally.

1. Introduction

In plant disease management, early detection surveillance is the process of searching a population to determine whether or not an invasive pathogen is present. The aim is to discover the invader before it has reached high prevalence so that a programme of control or containment can be instigated with as little as possible cost. Control of an invasive species is difficult and eradication less feasible if the invading population is not detected at an early stage [1,2]. For example, median losses from a potential Foot and Mouth Disease epidemic in California have been predicted to rise from $2.3 to $69 billion if the delay in detection of the index case increases from 7 to 22 days [3]. For plant diseases, the spatial scale of the population under surveillance will vary depending on the remit of the land manager, from the monitoring of a small woodland or farm conducted by a landowner to a national scale programme conducted by a regulatory agency. In any case, the first problem faced by a practitioner is how much resource to commit to a particular surveillance programme (e.g. how many sites should be sampled and how often). This is a multi-faceted problem but should be underpinned by a quantitative understanding of what a given surveillance effort will actually deliver, i.e. the prevalence an invader will have reached when it is first discovered.

How surveillance effort relates to the probability to achieve early detection is not known and plant health managers are faced with a general lack of
guidance on what constitutes an appropriate level of surveillance effort. Consequently, surveillance effort is often not sufficient and detection of many emerging diseases occurs only by chance once the epidemic has reached a high prevalence in the environment, by when effective control is no longer realistic. A pertinent example is ash dieback (causal agent *Chalara fraxinea*) in the UK. Despite its well-known steady advance across Europe, the disease was found to be already widespread in the natural environment when it was first discovered in the UK in 2012 [4]. Conversely, it is also possible that too much resource may be committed to a surveillance programme, detracting funds from other important facets of the overall disease management effort. Previous studies have demonstrated how the allocation of surveillance effort can be determined by considering, for example, the spatial dynamics of the invading population [5], interactions with any preconceived management options [6] and the specific surveillance objective [7]. However, these studies do not answer the question: given a surveillance programme, what prevalence will an epidemic have reached when it is detected for the first time? Although detailed species-specific approaches are valuable, where there is sufficient time and information for adequate parametrization, a more fundamental understanding of how the dynamics of an invader and surveillance programme relate is still missing.

Modelling offers an opportunity to explore and quantify the performance of a disease surveillance programme in relation to an epidemic [8]. Parnell et al. [9] used a simple deterministic and non-spatial epidemic model to show that the prevalence an epidemic will have reached when it is first discovered (hereafter termed ‘discovery-prevalence’) can be determined using basic information on the epidemic and surveillance programme. The epidemic model used made several simplifying assumptions, including logistic growth of the epidemic and homogeneous dispersal of inoculum. If robust, this ‘rule of thumb’ would represent a novel and useful insight to guide decisions on what level of surveillance resource effort is appropriate in practice, i.e. to ensure an epidemic is still at low prevalence when it is first discovered. Subsequently, we test the performance of our rule of thumb method on realistic systems. To do this, we begin by testing it against an individual-based epidemic simulation model which incorporates demographic stochasticity and spatial spread of the epidemic. This is in contrast to the original model by Parnell et al. [9] which was non-spatial and deterministic. In addition, we conduct a sensitivity analysis on the spatial and stochastic epidemic model to test the circumstances where the rule of thumb is most and least accurate and to explore its generality. We then go further and ask if the predictive ability of the rule of thumb is also accurate when tested on observed data from real epidemics.

In the following sections, first we describe the simple epidemic model approach and the ‘rule of thumb’ approximation to calculate the prevalence an invading population will have reached when first detected (i.e. ‘discov-ery-prevalence’). We then describe and present results to test the approximation against: (i) the realistic individual-based simulation model, and finally against (ii) datasets describing the spatial and temporal spread of real observed epidemics. This is achieved by simulating survey programmes to determine detection-prevalence on the underlying epidemic spread data based on the simulated and observed datasets. We use the example of two bacterial pathogens of citrus which represent two of the most pressing emerging plant health problems worldwide; citrus greening (syn. Huanglongbing, HLB; causal agent *Candidatus Liberibacter*) and citrus canker (causal agent *Xanthomonas citri* subsp. *citri*). HLB is one of the most damaging tree diseases worldwide [10] and is spread by the psyllid vector *Diaphorina citri*. It is a particularly grave threat to citrus production in the Americas where it continues to spread to new regions [11]. Citrus canker is a bacterial pathogen spread by splash dispersal and wind-blown rain and is present in over 30 countries worldwide [12] with enormous socio-economic impact [13]. It was the subject of a $1 billion eradication programme in the USA, one of the largest attempts to eradicate a plant disease [14]. The economic significance of these pathogens, as well the presence of detailed knowledge of their spatial and temporally complex epidemiology and the availability of high-resolution epidemic data, make these systems ideal case studies to test the method.

2. Discovery-prevalence: a simple rule of thumb

The scenario we consider is a surveillance programme where a fixed number of observations, or samples, *N* are made at random from a homogeneous host population at regular intervals *Δ*. At some unknown point in time, the pathogen invades and the disease prevalence *q*(*t*) begins to increase logistically in the host population with growth rate *r* and initial level of infection *q*0:

\[
q(t) = \frac{1}{1 + ((1-q_0)/q_0)\exp(-rt)}. \tag{2.1}
\]

The disease is only detectable when it reaches a certain prevalence, *q*0, which is set at some threshold. The goal of the surveillance programme is to determine, for a fixed number of observations, or samples, *N*, the moment at which the epidemic is first discovered, *t*1. Finally, using equation (2.1) we transform *t*0 to *q* to give us the prevalence the epidemic will have reached when first discovered, *q*0.

If the epidemic invades at time *t*0 then the probability to detect an infected host plant at time *t*1 is given by the complement of the zeroth term of the binomial distribution \((1 - (1 - q(t_1))^k)\). The probability not to have detected the epidemic during all previous sampling rounds is calculated from \(\prod_{k=1}^{n} P(k) = (1 - q(t_{k-1} - i\Delta))^N\), where *k* is the total number of sampling rounds between *t*0 and *t*1 and depends on *Δ*. Multiplying these terms gives the probability to detect at time *t*1 conditioned on the epidemic invasion time *t*0 and the...
probability of not being detected in all sampling rounds before $t_1$:

$$P(t_1|t_0) = \prod_{i=1}^{k} (1 - q(t_1 - t_0))^N.$$  

(2.2)

To calculate $P(q^*|t_1)$, that is, the probability that the epidemic reached the prevalence $q^*$ given that epidemic was detected at $t_1$ from $P(t_1|t_0)$, we use the methodology by Parnell et al. [9]. Using Bayes theorem and assuming that invasion is equally likely at any time, we can say that

$$P(t_0|t_1) = \frac{P(t_1|t_0) \cdot \text{Pr}(t_0)}{\int_0^\infty P(t_1|t_0) \cdot \text{Pr}(t_0) \, dt_0}.$$  

(2.3)

Assuming epidemic dynamics are deterministic, the probability of the epidemic reaching prevalence $q^*$ given that it has been discovered at $t_1$ is obtained by performing a random variable transformation of $t_0$ using equation (2.1) [9]. For this we use that $t = t - t_0$ and assume that the prevalence and epidemic growth rate are low at the beginning of the epidemic. Using the Jacobian of the transformation [16], the probability of detecting the epidemic at prevalence $q^*$ given that it was detected at $t_1$ is

$$P(q^*|t_1) = \frac{1}{N} \cdot P(t_0|t_1).$$  

(2.4)

(a) The simple ‘rule of thumb’

Parnell et al. [9] demonstrated that, by assuming exponential increase of the epidemic and continuous sampling, the following approximation to equation (2.4) could be derived:

$$P(q^*|t_1) \approx \frac{N}{r \Delta} \exp \left( \frac{N}{r \Delta} q^* \right).$$  

(2.5)

The above equation provides an expression for the approximate distribution of the discovery-prevalence $q^*$ which is an exponential distribution. It follows that the mean discovery-prevalence $E(q^*)$ of $P(q^*|t_1)$ can be written as

$$E(q^*) = \frac{r \Delta}{N} = \frac{r}{N/r \Delta}.$$  

(2.6)

We term the expression in the above equation our ‘rule of thumb’ to calculate discovery-prevalence. The D% confidence interval $[A, B]$ is calculated from

$$\int_{A}^{B} P(q^*|t_1) \, dq^* = 1 - \frac{1}{2}(1 - D),$$  

(2.7)

and

$$\int_{0}^{\infty} P(q^*|t_1) \, dq^* P = \frac{1}{2}(1 - D),$$  

(2.8)

which gives

$$A = -\frac{r \Delta}{N} \ln \left( 1 - \frac{D}{2} \right)$$  

(2.9)

and

$$B = -\frac{r \Delta}{N} \ln \left( 1 - \frac{D}{2} \right).$$  

(2.10)

Thus, we can say that with probability $D/100$ the discovery-prevalence $q^*$ is

$$-\frac{r \Delta}{N} \ln \left( 1 - \frac{D}{2} \right) < q^* < -\frac{r \Delta}{N} \ln \left( 1 - \frac{D}{2} \right).$$  

(2.11)

The upper bound of the confidence interval corresponds to the upper percentile, i.e. the percentage of cases that will not exceed this discovery-prevalence. Equation (2.11) tells us that as $r$ increases the confidence interval also becomes larger, decreasing the accuracy that the expected discovery-prevalence gives for a disease.

3. Testing the rule of thumb on realistic systems

In the following sections, we test the rule of thumb on realistic systems. Testing the accuracy of the rule of thumb is a two-step process. First, we calculate an estimate of discovery-prevalence from the approximation (equation (2.6)). For this, we need an estimate of the epidemic growth rate $r$ and the parameters of our surveillance programme (sample size $N$ and sample frequency $\Delta$). Second, we simulate a surveillance programme, using the same sample size $N$ and sample frequency $\Delta$ as for the approximation, on spatial–temporal data from either: (i) a spatially explicit simulation model (as we do for the first example of HLB), or (ii) an observed real epidemic (as we do for the second example of citrus canker). We can then compare the estimate from the approximation rule of thumb (equation (2.6)) with the ‘real’ discovery-prevalence from the simulated surveillance programme on the spatial–temporal data. To test accuracy, we must recourse to using a simulation approach to determine discovery-prevalence since actual discovery-prevalences from real epidemics under real surveillance programmes are not easily obtained in practice, i.e. this would require a census of the entire host population immediately following first detection.

(a) Does the rule of thumb hold for a realistic epidemic model? Including space and stochasticity

(i) The epidemic simulation model

The model is spatially explicit in that we keep track of the spatial coordinates of each individual host unit and account for distance-dependent transmission processes. We use an SIR model formulation [17,18] whereby a host can be either susceptible ($S$), infected ($I$) or removed ($R$). The rate of transition of individual susceptible host $S$ to infected $I$ is a function of the distance of that host to other hosts that are already in the $I$ state:

$$\lambda_i(t) = \sum_{j \in I} \frac{\alpha_i}{2\pi \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}.$$  

(3.1)

The (normalized) dispersal kernel $\exp(-ad_{ij})$ describes the probability that an infectious host $j$ transmits inoculum and causes infection of susceptible host $i$ at distance $d$. This is a negative exponential function with scale parameter $\alpha$. Following infection the fraction of each host that displays detectable symptoms increases logistically over time (deterministically) with initial amount infected $s_0$ and rate of increase $r_s$. The transition from $S$ to $I$ is stochastic and simulated using the Gillespie direct algorithm [17].

(ii) Model parametrization

A model of the above form captures the spread patterns of a wide range of invasive plant pathogens where dispersal is distance-dependent (e.g. aerially transmitted fungal, bacterial
and vectored diseases) and similar models have also been used to describe invasive animal diseases [19] and invading insect pests [20]. Here, the model is parametrized for the invasion and spread of an HLB epidemic in commercial citrus plantings. Typically, plantings consist of a regular block of 1540 trees arranged in 14 rows of 110 trees where the spacing between rows is 7.5 m and the spacing between trees in the same row is 3.5 m. We therefore used this topology in all simulations.

Observations on the temporal disease progress of HLB-infected plantings and symptoms from Bassanezi et al. [21] were used for parametrization. The age of a citrus planting is known to be a key determinant of HLB disease progression [11]. For each age class, Bassanezi et al. [21] describe both the rate of symptom development over time in individual trees and the rate of increase in the number of infected trees within a grove. The assumption made is that the latent period is negligible and that symptoms begin to increase immediately following infection. We combined these to determine the overall exponential rate of symptom development \( r \) in the early stage of an epidemic within a planting consisting of each of the four age classes. Selecting parameters in our simulation to match these rates was the target of our parametrization procedure.

In particular, we needed to select values of transmission rate \( b \) and dispersal scale \( a \) parameters, on an age class by age class basis. In all cases we used a dispersal scale of 8 m (i.e. \( a = 1/8 \text{ m}^{-2} \)) for our dispersal kernel this is supported by observations from artificial release experiments [22]. We calibrated the transmission rate \( b \) via a line search. In particular, we performed replicates of the runs of the simulation model for each of a range of values of the infection rate \( b \) (1000 runs per value of \( b \)). The initial rate of increase of symptoms in the early part of each simulated epidemic (i.e. \( r \)) was calculated by fitting an exponential function to the severity in the simulated epidemic (figure 1a). For each age class, we then simply selected the value of \( b \) that gave the closest match to the rates reported by Bassanezi et al. [21] (see the electronic supplementary material, S4).

(iii) Determining discovery-prevalence
A set of epidemic simulations were run for each of the four tree age classes and a surveillance programme was simulated by randomly selecting \( N \) trees independently at regular intervals \( \Delta \). The time of the first round of sampling was randomly drawn uniformly from the interval \([0, \Delta]\) to allow random invasion time at \( t = 0 \) relative to the surveillance programme. The probability to detect any symptomatic tree encountered on a single survey was equal to the fraction of symptoms displayed by the tree at the time of sampling. When a detection-event occurred the simulation was stopped and the prevalence of symptoms in the planting was recorded, i.e. the discovery-prevalence. Two hundred simulations were performed each for a range of sample sizes and sampling intervals. The test of the rule of thumb method was then performed by comparing the resulting, simulated, discovery-prevalences directly to those predicted by the rule of thumb approximation (equation (2.6)) using the same epidemic growth rate.

In addition to simple random sampling, we also consider more spatially realistic sampling schemes which are typically sought in practice to make sample collection logistically feasible. In the case of commercial citrus plantings, this is usually done by sampling every \( N \)th tree, i.e. an inspector walking up and down rows of citrus trees in a planting and inspecting or sampling every \( N \)th tree, starting at a randomly chosen tree. We simulate this process in the epidemic simulation model for HLB.

(iv) Sensitivity analysis
The baseline values of the dispersal scale \( a \) and the infection rate \( b \) are varied in a sensitivity analysis to explore the circumstances, where the approximation is less or more accurate and the extent to which extrapolations can be made to other systems. For both \( a \) and \( b \), we performed a one-way sensitivity analysis, in which one parameter was altered with the other held at a constant value. Since we are interested in isolating the effect of the parameter change on the relative performance of the approximation, we recalibrated the value of \( r \) for each pair of parameters we tested before calculating the expected detection-prevalence from our rule of thumb.

(v) Results from the epidemic model and sensitivity analysis
The mean epidemic growth rates (per day) for each planting age class were 0.0143 (\( \beta \) 0.5), 0.0065 (\( \beta \) 0.23), 0.003 (\( \beta \) 0.11) and 0.0021 (\( \beta \) 0.08) for the age classes 0–2 years, 2–5 years, 6–10 years and more than 10 years (\( R^2 \geq 0.98 \) in each case) (see the electronic supplementary material, S4). The rule of

![Figure 1](http://rspb.royalsocietypublishing.org/)

Figure 1. (a) HLB disease progress curves illustrating the change in prevalence with time for the growth rates obtained from the exponential fit to each age class (dashed lines) and the mean from the full epidemic simulation model (solid lines). The different planting age classes are: age 0–2 years (black), 3–5 years (red), 6–10 years (green) and more than 10 years (blue). The exponential fitted lines are used to inform the ‘rule of thumb’ for discovery-prevalence and are compared for accuracy with the discovery-prevalences obtained from the full epidemic simulation model. (b) Example probability distribution for discovery-prevalence for HLB planting age class 0–2 years. The mean (equation (2.6)) and 95th percentile discovery-prevalences are shown for the rule of thumb (solid and open circles, respectively) and the HLB simulation model (dashed and dotted lines, respectively). The sample size \( N \) is 50 trees and sample interval \( \Delta \) 50 days. Results for all age classes are provided in the electronic supplementary material, S1.

The mean epidemic growth rates (per day) for each planting age class were 0.0143 (\( \beta \) 0.5), 0.0065 (\( \beta \) 0.23), 0.003 (\( \beta \) 0.11) and 0.0021 (\( \beta \) 0.08) for the age classes 0–2 years, 2–5 years, 6–10 years and more than 10 years (\( R^2 \geq 0.98 \) in each case) (see the electronic supplementary material, S4). The rule of
thumb approximation closely matched the discovery-prevalences from the spatially explicit stochastic simulation model for HLB for both the mean and 95th percentile (figures 1b and 2). In general, the higher the sample size and the lower the interval between samples, the lower the discovery-prevalence and the closer the match between the simulated and rule of thumb (equation (2.6)) discovery-prevalences (figure 2). Losses in accuracy were owing to overestimation of discovery-prevalence (figure 2). Simulated ‘systematic’ survey patterns (sampling every Nth tree in a citrus planting) resulted in similar mean discovery-prevalences to those obtained from simple random sampling and were also well within the 90% confidence intervals of the rule of thumb equation (figure 2). The sensitivity analysis revealed that the accuracy of the rule of thumb increased with increasing dispersal scale (1/a) and decreased with transmission rate β (figure 3). We also performed a sensitivity analysis on the 95th percentiles which revealed very similar patterns to the mean, i.e. reductions in accuracy were owing to an overestimate of the 95th percentile and accuracy increased with dispersal scale and decreased with transmission rate (see the electronic supplementary material, S1).

(b) Can we make the step to real epidemics?

Data directly observed from real systems introduce further biological and environmental complexities that determine epidemic dynamics and will not be fully captured by any
model. Therefore, we also ask how well our simple approximation works directly on data from a real epidemic system. We have access to a high-quality spatially and temporally resolved dataset from Florida concerning the invasion and spread of a disease of citrus trees (see the electronic supplementary material, S2). This makes it possible to test the effect of different surveillance programmes on a real system. In the below sections, we summarize the datasets and the methods used to calculate discovery-prevalences for a range of surveillance programme designs.

(i) Data collection: citrus canker disease in urban Miami
Residential citrus trees grown in urban areas are a major source of inoculum and point of entry for the disease into commercial citrus areas. As such they are a major focus for surveillance efforts. Having previously invaded and been eradicated on a number of occasions in Florida in the 1900s, a new citrus canker epidemic began in 1995. The dataset gathered to study the 1990s epidemic was described by Gottwald et al. [23], where the reader is referred for further information. The study sites ranged between 2.6 and 15.5 km² of urban area and contained a total of over 18 000 dooryard trees (trees located in the gardens of residential properties) with the spatial location of each tree recorded using GPS (electronic supplementary material, S2). Surveys were conducted at approximately 60 day intervals with at least three surveys in each study site. During each survey, all citrus trees were visually assessed for disease symptoms and initial dates of infection were estimated by dating individual lesions based on their phenotypic characteristics. Epidemics lasted approximately 540 days after which a period of sustained drought rendered trees non-susceptible to the bacteria and no further infections were reported during the duration of the study.

(ii) Determining discovery-prevalence
To test the accuracy of the rule of thumb, we first determined discovery-prevalence from the approximation (equation (2.6)). To do thus we need to determine an appropriate estimate of the epidemic growth rate ($r$). This was done by fitting a logistic function to the temporal citrus canker progress data (see the electronic supplementary material, S2). Once we had determined the discovery-prevalence from the approximation we then simulated a surveillance programme on the real data to directly compare with this. (Note that since it is important to separate trial and test data, for validation purposes, each site was assigned an epidemic growth rate, $r$, for the approximation (equation (2.6)) from a different randomly assigned site than was used for the simulated surveillance programme on the real data.) Monte Carlo
simulations were performed to simulate a disease surveillance programme in each of the study sites separately. A number of trees $N$ were selected at random at regular intervals $\Delta$, with a random starting point chosen on the interval $[0, \Delta]$; this was possible since the data are resolved to a daily time step as infection dates were estimated in the field when the data were collected. If an infected tree was selected, then the simulation was stopped and the prevalence of the epidemic recorded. One thousand runs of the simulation were performed on each study site producing a probability distribution of discovery-prevalences. This was repeated for different combinations of sample size $N$ and sampling interval $\Delta$ and used to determine discovery-prevalences for a range of surveillance programmes based on the rule of thumb (equation (2.6)) and compared with the discovery-prevalences from the simulation on the observed data. Note that since only the infection status of each tree was recorded (infected or not infected), and not the fraction of each tree that was symptomatic, here ‘prevalence’ is approximated by the fraction of infected trees in the population rather than the average severity in the population as in equation (2.6) and the HLB simulation model.

As was done for the HLB epidemic model, in addition to simple random sampling, we also consider more practical sampling plans. Owing to the spatially heterogeneous nature of the citrus trees in residential areas, sampling is usually systematized by performing stratified random sampling. For the datasets from Miami, we simulated this by separating the host population into regular rectangular strata and one randomly selected tree is sampled per strata. Thus, for a sample size of one the area would consist of a single stratum, for a sample size of two the area is split into two rectangular strata and so on. This type of stratification is regular practice in urban areas and United States Department of Agriculture, Animal and Plant Health Inspection Service used this method to monitor urban areas in Florida for citrus canker invasion into new areas during the epidemic in the 1990s as part of its Sentinel Tree Program [12].

(iii) Results from the real epidemic data
The epidemic growth rates (per day) for each study site from Miami 1 to 4 were 0.0157 ($R^2 = 0.987$), 0.0138 ($R^2 = 0.996$), 0.0171 ($R^2 = 0.993$) and 0.0163 ($R^2 = 0.995$), respectively. To test the accuracy of the rule of thumb, each site, Miami 1–4, was randomly assigned the following growth rates for input into the rule of thumb (equation (2.6)) 0.0138, 0.0157, 0.0163 and 0.0171, respectively. Here we show results for a single site (figure 4). Similar results were found across sites and full results are available in the electronic supplementary material, S2.

The discovery-prevalences simulated from the observed citrus canker epidemic data were closely approximated by the rule of thumb (figure 4; electronic supplementary material, S2). Simulated discovery-prevalences were well within the 90% confidence intervals associated with our rule of thumb (figure 4; electronic supplementary material, S2). The accuracy of the rule of thumb generally increased with sample size, $N$, and decreased with sampling interval, $\Delta$ (figure 4) as was also found for the HLB model (figure 2). Stratified random sampling was similarly closely predicted by the rule of thumb approximation (figure 4) and tended to result in discovery-prevalences marginally higher than for simple random sampling (figure 4).

4. Discussion
We have introduced a simple rule of thumb to estimate the prevalence an epidemic will have reached when first discovered (i.e. discovery-prevalence) by a surveillance programme. We have tested the rule of thumb on simulated and observed epidemics representing two of the most destructive pathogens currently faced by plant health officials worldwide, citrus canker and citrus greening (syn. HLB). The accuracy of the rule of thumb in predicting the discovery-prevalences of these spatially and temporally complex epidemics is surprising given its simplicity. What this highlights is that, although complex spatial and temporal patterns emerge as an epidemic develops, for the purposes of early detection only the initial phase of the population expansion is relevant.
Advantageously, this initial phase is characterized by much simpler spatial and temporal patterns. Our simple rule of thumb can thus provide useful insights into the outcome of early warning surveillance in apparently complex systems.

We have demonstrated the robustness of the rule of thumb to a range of epidemiological and sampling parameters. Cases where the accuracy is diminished can be largely attributed to violations of the key assumption of exponential increase that is made in the derivation of the approximation (see the electronic supplementary material, S1 for additional discussion). The accuracy of the rule of thumb decreases with decreasing sampling effort, i.e. decreasing sample size, N and increasing sampling interval Δ (figures 2 and 4). With lower sampling effort, the epidemic is discovered later at which point it may have already left the exponential phase of expansion. The exponential increase assumption is also violated, and thus accuracy diminishes, for cases where dispersal scale is particularly limited (i.e. short range population spread; figure 3a–d) and where transmission rate is high (figure 3e–h). Note that for the citrus canker example data were only available on the number of infected trees and not the fraction of each that was infected. This may have influenced the accuracy of the rule of thumb, which assumes our default definition of prevalence encompassing both the fraction of each host infected as well as the number of individuals infected. We also assume that all infected trees are detectable and thus that inspection efficiency per tree is high. Citrus canker produces visible lesions on the surface of plant material and therefore trees are typically detectable shortly after infection. However, depending on the inspection methods and efficiency of a particular survey, this assumption may not hold and should be considered as it could lead to underestimates of discovery-prevalence.

Exponential increase is a coarse approximation for all but the simplest of invading populations. However, for any particular system, the accuracy and usefulness of our approach will hinge on in how far this assumption is violated, at low prevalence, for the particular invader of interest, and the intensity of the surveillance programme. Although the method is not perfect, the accuracy achieved based on little information arms practitioners with a useful generic tool to inform decisions about the level of resource to commit to any particular early detection surveillance programme. Standard statistical methods to estimate detection probabilities are available, most of these are based on binomial sampling theory [24,25]. However, unlike our method, such approaches do not include any aspects of the epidemiology of the invading population. These approaches assume that epidemic prevalence is constant and often require that an assumption of ‘true prevalence’ be made. In reality, epidemics invade at an unknown time, making estimates of true prevalence difficult at any single point in time. We offer an alternative approach that accounts for multiple rounds of sampling, unknown invasion time and the post-invasion rate of increase of the invading population.

The use of our approach is in its simplicity and it hinges only on an estimate of the epidemic growth rate, r. The accuracy of this estimate depends on the availability of data in the host population of concern and the level of fit achieved. The ideal situation will occur when it can be accurately estimated from observed epidemics in similar host populations. For example, when an epidemic is advancing on a region and data to estimate growth rates from epidemics in neighbouring regions are available. For example, it may have been possible to use data on the rate of epidemic growth of ash dieback from the continent to inform surveillance efforts in the UK as the epidemic spread westward from Poland [26]. In the absence of suitable data, estimates of growth rates can be coarsely estimated from pathogens which share similar dispersal characteristics. Where no information is available the rule of thumb provides a tool to explore best and worst case scenarios and so remains a useful component of the decision-making process in such circumstances.

It should be noted that here we assume that the definition of prevalence relates to both the epidemic growth rate (the increase in the fraction of the host population infected over time) and the discovery-prevalence, since we assume prevalence is symptomatic, i.e. detectable prevalence. In the current paper, we do not consider asymptomatic infection but a useful extension of the method would be to include a lag time for asymptomatic to symptomatic infection, for example, if the discovery-prevalence required refers to the prevalence of asymptomatic infection but the detection method is assessment of symptomatic tissue. Similarly, an analysis of the effect of a false-negative and false-positive rates could be factored into the original calculation of the probability to detect an epidemic (equation (2.2)) and would yield insights into the effect of sample sensitivity and specificity. It should be noted that the method assumes that surveillance has begun before the epidemic has invaded, rather than simply before it has been detected. If an invasion has already begun when surveillance is initiated, then the method will underestimate discovery-prevalence.

In summary, the rule of thumb presented in this paper offers a convenient and insightful tool that can be used by plant health practitioners to motivate the choice of surveillance effort for a potential invader. The method assumes random sampling and therefore will serve as an upper bound to more effective targeted sampling strategies such as those based on epidemic risk [27]. Although we have validated the rule of thumb on two plant diseases and explored its sensitivity to key plant disease epidemiological parameters, its simplicity lends itself to exploration for a broad range of potential invasive species and infectious disease applications. Many invasive species and infectious diseases are, in their early phases, well characterized by the assumptions made by our approach, e.g. exponential increase. Any new application would need to be tested and validated and the approach used in this paper could be used as a blueprint to do so.

Data accessibility. Data from this study are not available.

Authors’ contributions. S.P., F.V.B., N.J.C. and T.R.G. designed the study. S.P., N.J.C. and V.A.C. conducted the analysis. S.P. wrote the manuscript. All authors contributed to data interpretation, manuscript editing and discussion.

Competing interests. We have no competing interests.

Funding. S.P., T.R.G. and F.V.B. received funding from the US Department of Agriculture Farm Bill. S.P., F.V.B. and V.A.C. received support from the U.K. Biotechnology and Biological Sciences Research Council (BBSRC) through Rothamsted Research. F.V.B. received funding from the Bill and Melinda Gates Foundation.
References


